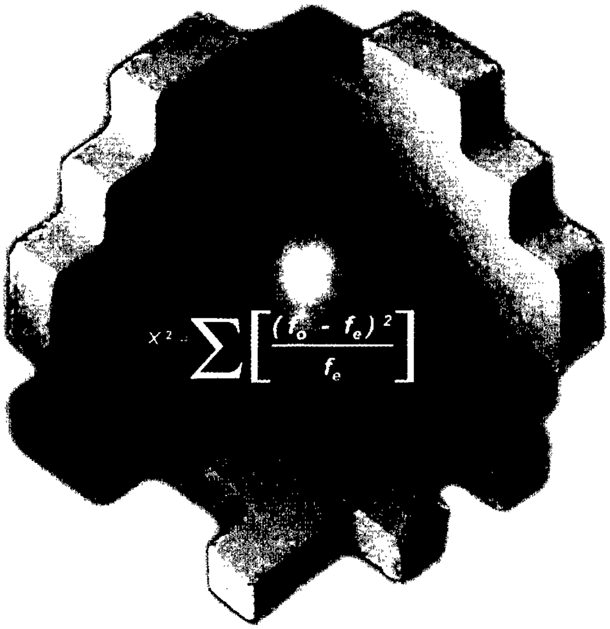
Defense Manpower Data Center



Evaluating Large-Scale Training Simulations

Volume I: Reference Manual

Henry Simpson

Henry Simpson

Evaluating Large-Scale Training Simulations

Volume I: Reference Manual

Defense Manpower Data Center · Seaside, California

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
December 1999

3. REPORT TYPE AND DATES COVERED
Final

4. TITLE AND SUBTITLE
EVALUATING LARGE-SCALE TRAINING SIMULATIONS, VOLUME I: REFERENCE MANUAL

5. FUNDING NUMBERS

6. AUTHOR(S)
Henry Simpson

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Defense Manpower Data Center
DoD Center, Monterey Bay
400 Gigling Road
Seaside, CA 93955-6771

8. PERFORMING ORGANIZATION REPORT NUMBER
DMDC Technical Report 99-05

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Deputy Under Secretary of Defense (Readiness)
4000 Defense, The Pentagon
Washington, DC 20301-4000

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

Objectives of the manual are to (1) provide guidance to help analysts design meaningful training effectiveness evaluations, (2) describe procedures for alternative methods of conducting training effectiveness evaluations, and (3) provide examples of training effectiveness evaluations that may be used as models to emulate. Chapter 1 (Introduction) describes the problem and issues, objectives, and method. Chapter 2 (Building an Evaluation Framework) explains why people conduct evaluations. Chapter 3 (Evaluation Methods) describes evaluation methods and provides examples of their application. Chapter 4 (Case Studies) describes well-documented evaluations: SIMNET/CCTT (Simulation Networking/Close Combat Tactical Trainer) and MDT2 (Multi-service Distributed Training Testbed). Chapter 5 (Evaluation Problem Areas) contrasts laboratory and field evaluations, discusses lessons learned from past evaluations, and critiques field evaluation practice. Chapter 6 (Procedural Guidance) identifies and summarizes published evaluation guidance. Chapter 7 (Evaluation Criteria) discusses how evaluation criteria differ depending upon evaluation method, for small- and large-scale evaluations, and depending upon evaluation perspective (training versus system developer versus modeling and simulation). Chapter 8 (Evaluation Framework) presents the evaluation framework in terms of evaluation principles and a description of the timing of evaluation events, their purpose, and relevant dependent variables—linked to relevant examples and procedural guidance.

14. SUBJECT TERMS
large-scale training simulation, training evaluation, evaluation methods, training effectiveness, military training

15. NUMBER OF PAGES
190

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Unlimited

To Jesse Orlansky

Preface

The Office of the Inspector General, Department of Defense, recommended that the Under Secretary of Defense for Personnel and Readiness establish policy and procedures to evaluate the training and cost-effectiveness of large-scale training simulations. One of the Under Secretary's responses to this request was to direct the Defense Manpower Data Center to develop guidelines to help evaluators conduct evaluations and to develop a historical training effectiveness data base. This volume describes the work performed by Defense Manpower Data Center in response to the Under Secretary's request and the resulting products and actions: Guidance to help evaluators design meaningful training effectiveness evaluations, descriptions of procedures for alternative methods, methodological examples, historical training effectiveness data base, and user access to the data base. These products and actions are intended to help the military Services determine when and how to evaluate the training and cost-effectiveness of large-scale training simulations.

This manual consists of two volumes: I (Reference Manual) and II (User's Manual). Volume II is designed to give readers a quick start introduction to training evaluation and a roadmap to the more in-depth content of Volume I. Readers are encouraged to start with Volume II.

The views expressed in this manual are those of the author, are not official, and do not necessarily reflect an official policy position of the Services, the Department of Defense, or the U.S. Government.

Acknowledgments

Dan Gardner sponsored the development of this manual. I thank him for his encouragement and patience.

More than two dozen individuals contributed ideas to this manual. The manual also uses many case studies based on 250 training evaluations conducted by literally hundreds of researchers. I have attempted to knit the ideas and evaluation practices into an overall evaluative framework for large-scale training simulations. The names of the many influences on this manual are too numerous to list here but will be evident from the citations.

Fred Hartman facilitated the review of early drafts of this manual. I thank him for providing many suggestions for improving the manual.

Thanks to Richard Kass for allowing me to reproduce his *Test Officer's Guide for Designing Valid Tests and Experiments* in Chapter 6.

Randy Marks of DMDC designed and composed the manual with the assistance of Aaron Toy, who created the illustration for the cover .

Dan Gardner, Fred Hartman, Jack Leather, and John Morrison reviewed this manual. Herbert Bell, David Bessemer, Edward George, Jack Hiller, Richard Kass, Douglas Macpherson, Angelo Mirabella, Frank Moses, Randy Oser, Robin Rose, and Uldi Shvern reviewed portions of the manual describing the evaluation framework. Organizational affiliations are shown below.

The individuals listed on the facing page provided information, opinions, documents, and in other ways supported the development of the manual. The presence of a name here or above does not indicate endorsement of the manual. I assume full responsibility for its contents and any errors, omissions, or oversights.

Henry Simpson
Seaside, California

Dee Andrews—Air Force Armstrong Laboratory
Herbert Bell—Air Force Armstrong Laboratory
David W. Bessemer—Army Research Institute
John Boldovici—Army Research Institute
Janis A. Cannon-Bowers—Naval Air Warfare Center Training Systems Division
Kenneth D. Cross—Bayview Research
Philip Djang—TRADOC Analysis Center
John Ellis—Navy Personnel Research and Development Center
Dorothy Finley—Army Research Institute
Dan Gardner—Office of the Deputy Under Secretary of Defense for Readiness
Edward L.George—TRADOC Analysis Center
Joseph Hagman—Army Research Institute
Fred Hartman—Institute for Defense Analyses
John Hayes—Army Research Institute
Jack H. Hiller—Army Research Institute
Donald Johnson—Office of the Deputy Under Secretary of Defense for Readiness
Richard Kass—Army Test and Experimentation Command
Peter Kincaid—University of Central Florida Institute for Simulation and Training
Richard Laferriere—TRADOC Analysis Center
Jack Leather—Defense Manpower Data Center
Douglas Macpherson—Army Research Institute
Angelo Mirabella—Army Research Institute
Franklin L. Moses—Army Research Institute
Randy Oser—Naval Air Warfare Center Training Systems Division
Jesse Orlansky[1]—Institute for Defense Analyses
Ruth Phelps—Army Research Institute
William Rankin—Naval Air Warfare Center Training Systems Division
J. Wesley Regian—Air Force Armstrong Laboratory
Robin Rose—TRADOC Analysis Center
Eduardo Salas—Naval Air Warfare Center Training Systems Division
Robert Seidel—Army Research Institute
Uldi Shvern—Army Operational Test and Experimentation Command
Henry L. Taylor—Institute of Aviation, University of Illinois
Diana Tierney—Deputy Chief of Staff for Training, TRADOC
Robert Worley—Institute for Defense Analyses

[1] Deceased.

EXECUTIVE SUMMARY

Problem and Issues

The Department of Defense Office of the Inspector General (DoDIG) conducted an audit concerning the impact on readiness of training simulators and devices. The audit focused particular attention on shortcomings in evaluation of large-scale training simulations (LSTS). The DoDIG recommended that the Under Secretary of Defense for Personnel and Readiness establish policy and procedures to evaluate the training and cost-effectiveness of LSTS. One of the Under Secretary's responses to this request was to direct the Defense Manpower Data Center (DMDC) to develop guidelines to help evaluators conduct evaluations and a historical training effectiveness data base. This manual describes the work performed by DMDC and the resulting products and actions: Guidance to help evaluators design meaningful training effectiveness evaluations, descriptions of procedures for alternative methods, methodological examples, historical training effectiveness data base, and user access to the data base. These products and actions are intended to help the military Services determine when and how to evaluate the training and cost-effectiveness of LSTS.

Objectives

Objectives of this manual are to:

- Provide guidance to help analysts design meaningful training effectiveness evaluations.
- Describe procedures for alternative methods of conducting training effectiveness evaluations.
- Provide examples of training effectiveness evaluations that may be used as models to emulate.

Method

The method consisted of literature review, development of a historical training and cost-effectiveness data base, analyses, development of guidelines, identification of case studies, and review of findings by subject-matter experts.

Evaluation Guidance

This manual contains evaluation guidance in eight chapters and two appendices. Chapter 1 (Introduction) describes the problem and issues, objectives, method, and shows where the manual addresses its objectives. Chapter 2 (Building an Evaluation Framework) explains why people conduct training effectiveness evaluations and starts to build an evaluation framework for LSTS by attempting to answer fundamental questions about the evaluation process (Whose training is evaluated? What is evaluated? Where to evaluate? How to evaluate? What are evaluation criteria? When to evaluate?). Chapter 3 (Evaluation Methods) describes the methods commonly used in military training effectiveness evaluations and provides examples of their application. Chapter 4 (Case Studies) reviews two well-documented evaluations of LSTS: SIMNET/CCTT (Simulation Networking/Close Combat Tactical Trainer) and MDT2 (Multi-Service Distributed Training Testbed). Chapter 5 (Evaluation Problem Areas) contrasts laboratory and field evaluations, discusses lessons learned from past evaluations, and critiques field evaluation practice. Chapter 6 (Procedural Guidance) identifies and summarizes published evaluation guidance. Chapter 7 (Evaluation Criteria) discusses how evaluation criteria differ depending upon evaluation method, for small- and large-scale evaluations, and perspective. Chapter 8 (Evaluation Framework) presents the evaluation framework. Appendix A (Reference Lists for Chapter 3) contains reference lists. Appendix B (Acronyms) defines acronyms. Author and Subject indexes are provided.

CONTENTS

CHAPTER 1. INTRODUCTION 1

Overview 1
Problem and Issues 2
Objectives 4
Method 4
Road Map: Where This Guide Addresses Each of Its Objectives 7

CHAPTER 2. BUILDING AN EVALUATION FRAMEWORK 9

Why Evaluate? 10
Whose Training Is Evaluated? 12
What Is Evaluated? 14
Where To Evaluate? 16
How To Evaluate? 16
What Are Evaluation Criteria? 21
When To Evaluate? 22

CHAPTER 3. EVALUATION METHODS 25

Experiments 25
Judgment-Based Evaluations 36
Analytical Evaluations 40
Surveys 47

CHAPTER 4. CASE STUDIES 49

Finding Cases 49
Two Good Cases 50
SIMNET/CCTT 50
MDT2 58
Reference List 4-1. SIMNET/CCTT Publications 67
Reference List 4-2. MDT2 Publications 70

CHAPTER 5. EVALUATION PROBLEM AREAS 73

Field Versus Laboratory Evaluations 73
Lessons Learned 77
Critiques of Field Evaluation Practice 80

CHAPTER 6. PROCEDURAL GUIDANCE 85

Evaluation Methods 85
System and Program Evaluation Frameworks 93
Collective and Team Training 95
Reference List 6-1: Procedural Guidance 97

CHAPTER 7. EVALUATION CRITERIA 101

Defining Evaluation Criteria 101
Evaluation Criteria for an LSTS Evaluation 106

CHAPTER 8. EVALUATION FRAMEWORK 111

Evaluation Objectives 111
Evaluation Principles 113
Evaluation Framework 116
Reference List 8-1. Evaluation Framework 131

REFERENCES 137

APPENDIX A. REFERENCE LISTS FOR CHAPTER 3 155

Overview 155
Reference List A-1. True Experiment 155
Reference List A-2. Pre-Experiment 161
Reference List A-3. Quasi-Experiment 163
Reference List A-4. Test 164
Reference List A-5. Transfer Experiment 166
Reference List A-6. Ex Post Facto 167
Reference List A-7. Judgment (Users) 169
Reference List A-8. Judgment (SMEs) 170
Reference List A-9. Judgment (Analysts) 171
Reference List A-10. Analysis (Evalute) 172
Reference List A-11. Analysis (Compare) 174
Reference List A-12. Analysis (Optimize) 175
Reference List A-13. Survey 176

APPENDIX B. ACRONYMS 179

AUTHOR INDEX 183

SUBJECT INDEX 187

FIGURES

4-1 TARGETs data for planning phase (from Orlansky et al., 1997) 63
4-2 TOM data for planning phase (from Orlansky et al., 1997) 64
5-1 Notional representation of a simple inverse relationship between internal and external validity for laboratory and field studies (adapted from Shute and Regian, 1993). 76
6-1 Test Officer's Guide for Designing Valid Tests and Experiments (from Kass, 1997, reproduced courtesy of the author) 88

TABLES

2-1 Some Key Evaluation Questions 10
2-2 Training Content, Echelon, and Training Versus Education Taxonomy 13
2-3 Training Type and Subtype Taxonomy 15
2-4 Frequency of Usage of Common Evaluation Methods and Submethods 19
2-5 Levels of Data Commonly Associated with Evaluation Methods 21
3-1 Three True Experimental Designs (from Campbell & Stanley, 1966) 26
3-2 Three Pre-Experimental Designs (from Campbell & Stanley, 1966) 28
3-3 Ten Quasi-Experimental Designs (from Campbell & Stanley, 1966) 30
3-4 Three Classes of Transfer Designs (adapted from Pfeiffer & Browning, 1984) 34
3-5 Frequencies and Percentages of Usage of Analyst-, SME-, and User Judgment-Based Evaluations in TCEF Sample 38
3-6 Frequencies and Percentages of Usage of Three Classes of Analysis (Evaluate, Compare, or Optimize) for Existing Versus Hypothetical Systems 41

TABLES (continued)

4-1 SIMNET/CCTT Evaluations by Authors and Evaluation Method and Submethod......... 54

4-2 Simulators, Service, Types of Participant, and Locations Linked in MDT2 (adapted from Orlansky et al., 1997). 60

4-3 Dimensions and Subdimensions of TOM (Teamwork Observation Measure) (adapted from Orlansky et al., 1997). 62

4-4 Bombing Performance for February 1995 Exercise (from Orlansky et al., 1997)............... 64

4-5 CAS Kills, Misses, and Average Engagement Time for February 1995 Exercise (adapted from Orlansky et al., 1997). 64

4-6 Results of MDT2 Survey (from Orlansky et al., 1997).. 65

5-1 Potential Limitations of Reported Findings on Training Device Effectiveness (Based on N=35 evaluations) (adapted from Kraemer & Rowatt, 1993)............................. 83

7-1 Dependent Variables Used in SIMNET/CCTT Evaluations by Authors and Evaluation Methods 102

7-2 Consolidated List of Recommended Dependent Measures by Author.............................. 108

7-3 Consolidated List or Recommended Dependent Measures with Descriptions 108

8-1 Common Objectives for Conducting Training Effectiveness Evaluations........................ 112

8-2 Descriptive Summary of Representative Evaluations for Evaluation Objective 2A: Predict Training Effectiveness.. 118

8-3 Descriptive Summary of Representative Evaluations for Evaluation Objective 2B: Measure Training Effectiveness 119

8-4 Evaluation Criteria by Study for Objective 2B: Measure Training Effectiveness............. 119

8-5 Descriptive Summary of Representative Evaluations for Evaluation Objective 2C: Determine Training Effectiveness 121

8-6 Descriptive Summary of Representative Evaluations for Evaluation Objective 3B: Measure Transfer of Training.. 123

8-7 Evaluation Criteria by Study for Objective 3B: Measure Transfer of Training................. 123

8-8 Descriptive Summary of Representative Evaluations for Evaluation Objective 4B (Measure User Acceptance) and 4C (Determine User Acceptance).................................... 127

8-9 Descriptive Summary of Representative Evaluations for Evaluation Objective 5(AB): Support Training Design 128

8-10 Descriptive Summary of Representative Evaluations for Evaluation Objective 5C: Determine Training Status 129

1 INTRODUCTION

Overview

Chapter 1 (Introduction) describes the problem and issues, objectives, method, and shows where the manual addresses each of its objectives.

Chapter 2 (Building an Evaluation Framework) explains why people conduct training effectiveness evaluations and starts to build an evaluation framework for large-scale training simulations (LSTS) by attempting to answer some fundamental questions about the evaluation process (Whose training is evaluated? What is evaluated? Where to evaluate? How to evaluate? What are evaluation criteria? When to evaluate?).

Chapter 3 (Evaluation Methods) describes the evaluation methods commonly used in military training effectiveness evaluations and provides examples of their application.

Chapter 4 (Case Studies) describes two well-documented evaluations of LSTS: SIMNET/CCTT (Simulation Networking/Close Combat Tactical Trainer) and MDT2 (Multi-Service Distributed Training Testbed).

Chapter 5 (Evaluation Problem Areas) contrasts laboratory and field evaluations, discusses lessons learned from past evaluations, and critiques field evaluation practice.

Chapter 6 (Procedural Guidance) identifies and summarizes published evaluation guidance from a variety of sources.

Chapter 7 (Evaluation Criteria) discusses how evaluation criteria differ depending upon evaluation method, for small- and large-scale evaluations, and depending upon evaluation perspective (training versus system developer versus modeling and simulation).

Chapter 8 (Evaluation Framework) presents the evaluation framework in terms of evaluation objectives and principles. It also describes evaluation events and links them to relevant examples and procedural guidance.

All of the references cited in the chapters of this manual are listed in References.

The appendices contain information to support the discussion elsewhere in the manual and are referenced at appropriate locations.

Chapter 3 has 13 separate reference lists for the 250 evaluations it is based on. These lists appear in Appendix A (Reference Lists for Chapter 3). Chapters 4, 6, and 8 include separate reference lists to make it easier for readers to compile lists of works to obtain, should they so desire.

Appendix B (Acronyms) defines acronyms used in this manual.

Problem and Issues

The DoD Office of the Inspector General (DoDIG) conducted an audit concerning the impact on readiness of training simulators and devices (DoDIG, 1997). The audit focused attention on shortcomings in evaluation of LSTS. Large-scale training simulations are multi-million dollar simulations that may link together hundreds of participants at many different geographic locations to interact on a virtual battlefield.[2] Large-scale training simulations are more complex than traditional training devices (e.g., stand-alone gunnery simulators, flight simulators, and maintenance simulators) by an order of magnitude or more. Large-scale training simulations are classified as virtual, constructive, or advanced distributed simulations.[3] The Services have little experience evaluating them and there are no standard evaluation methods.

An LSTS has a total procurement cost of hundreds of millions of dollars. The DoDIG estimated that the overall acquisition cost of training systems by the Military Departments now exceeds $1.5 billion per year. This cost represents the cumulative cost of several different systems, each itself a multi-million dollar investment. The DoDIG had difficulty obtaining cost estimates for the systems it examined, although it estimated that the program cost for eight of them was approximately $2.6 billion.[4]

The DoDIG concluded that the "Military Departments have not demonstrated that large-scale computer training simulations being developed will be as effective as current training methods."

The Under Secretary of Defense for Personnel and Readiness (USD [P&R]) had conducted a review and begun to take actions to correct shortcomings. A 1995 USD (P&R)-sponsored review of the cost-effectiveness analysis of training (CEAT) in the DoD concluded, among other things, that methods for conducting cost- and training-effectiveness analyses were not well defined and that existing procedural guidance was inadequate (Simpson, 1995). The review recommended that a set of resources be developed to help evaluators: (1) guidelines to select the most suitable CEAT method based on circumstances, (2) procedural descriptions of methods, and (3) examples of completed studies linked to each method to use as case studies.

[2] The prototypical large-scale simulation is SIMNET (simulation networking), a product of technological developments of the early 1980s. SIMNET has never been definitively evaluated. Approximately three dozen studies dealing with one aspect or another of SIMNET's training effectiveness have been published. None is sufficiently comprehensive to settle conclusively the matter of SIMNET's training effectiveness. (These studies are reviewed in Chapter 4 [Case Studies] of this manual.) SIMNET is being superseded by the CCTT (close combat tactical trainer).

[3] A number of different and sometimes confusing classification schemes are used to describe simulations. The following definitions are adapted from Simpson, West, and Gleisner (1995). In defining simulations, it is useful first to consider simulation in terms of people and systems, and second whether the simulation represents both, neither, or either one. (A person who participates in a simulation is real, but other participants may be either real or simulated.) *Virtual simulation (VS)* involves real people interacting with simulated systems in a many-on-many environment. *Stand-alone, single-system simulation*—like VS, this involves real people interacting with simulated systems, but typically in a one-on-one environment. Examples are gunnery, crew, flight, operator, and maintenance simulators; alone, these are too small in scale to be classified as LSTS. *Live simulation* combines real people and real systems, generally in a many-on-many environment. *Constructive simulation* combines simulated combat forces and simulated systems in a computer-based model of combat in which combat systems are controlled by formal rules of movement, engagement, and casualty resolution. *Advanced Distributed Simulation (ADS)* is the combination of live, virtual, and constructive simulation.

[4] Department of Defense Office of the Inspector General (1997) estimated costs, by program, were Close Combat Tactical Trainer (CCTT), $846 million; WARSIM 2000 (Warfighter's Simulation), $172 million; Battle Force Tactical Trainer (BFTT), $165 million; Maritime Simulation (MARSIM), $142 million; Joint Tactical Combat Training System (JTCTS), $270 million; Synthetic Theater of War

In its audit report, the DoDIG recommended that the USD (P&R) establish policy and procedures for evaluating the training effectiveness and cost-effectiveness of LSTS (DoDIG, 1997). In response, the USD (P&R) committed to:

> ...developing policy and guidelines for conducting cost-effectiveness analyses of large-scale training simulations that: (1) allow analysts to select the best method under the circumstances, (2) describe the procedures for the various methods, and (3) provide examples that may be used as models to emulate. The USD (P&R) has also committed to establishing a historical training effectiveness data base and will ensure appropriate access to this information.... (Kaminski, P.G. [1997, March 17], p. 6).

These commitments echo the recommendations of the USD (P&R) review. In addition, they include requirements to establish a historical training effectiveness data base and access to that data base by training evaluators. They commit the USD (P&R) to the following products and actions:

- Policy (revisions as necessary)
- Guidelines to select best evaluation methods
- Descriptions of procedures for alternative methods
- Methodological examples
- Historical training effectiveness data base
- User access to data base

The present manual deals exclusively with training effectiveness evaluation. Studies that evaluate training are often referred to as TEAs (Training Effectiveness Analyses).[5] The cost-effectiveness analysis of training requires evaluators to conduct both TEAs and cost analyses and later to integrate the analyses. The mechanics for doing this are fairly well understood and described (e.g., Orlansky, 1985, 1989; Sassone and Schaffer, 1985; Simpson, 1995). Cost analysis is well defined; training effectiveness evaluation is not and is more difficult.

The use of LSTS is a recent development. They have come into widespread use in the last decade or so and the number of published evaluation studies is small. Many simulations are developed without publishing their training evaluation reports in a way that allows ready access; many study reports are not submitted to the Defense Technical Information Center for archiving. The paucity of LSTS evaluation studies is a problem for evaluators looking to the historical record for case studies or examples. Evaluators are forced to look beyond LSTS evaluations to training evaluation studies of training media, methods, programs, and small-scale simulators.[6]

Advanced Concept Technology Demonstration (STOW ACTD), $442 million; Joint Simulation System (core) (JSIMS), $154 million; Distributed Interactive Simulation (DIS), $500 million.

[5] TEA is also sometimes used as an acronym for Training Effectiveness Assessment. For purposes of this manual, the terms have the same meaning and are used interchangeably.

[6] Of the 250 studies used as the basis for analysis in this manual, approximately one-fourth (65) deal with LSTS. These 65 reports represent most of the published training evaluations on LSTS in the last 10 years.

Objectives

Objectives of this manual are to:

- Provide guidance to help analysts design meaningful training effectiveness evaluations.
- Describe procedures for alternative methods of conducting training effectiveness evaluations.
- Provide examples of training effectiveness evaluations that may be used as models to emulate.

Method

The method consisted of literature review, development of a historical training and cost-effectiveness data base, analyses, development of guidelines, identification of case studies, and review of findings by SMEs (Subject-Matter Expert).

Conduct Literature Review

A literature review was the main source of information in this manual. The review focused primarily on applied, non-theoretical studies conducted by or for the Services or DoD during the period 1974-1998, and weighted toward the most recent decade. The review also included relevant material published in the open literature, primarily research summaries and evaluation methodological guidance.

The review was used to determine alternative training evaluation methods, identify methodological examples, and build a historical training effectiveness data base. Subsequent analyses of documents enabled the development of training evaluation guidelines.

The review included four classes of documents: Evaluations, Research Summaries, Methods, and Policy. Each class includes one or more different types of documents:

- Evaluations are studies conducted to evaluate a form of training; for example, simulation, training medium, method, or program. Evaluations include documents such as cost analyses and cost-benefit analyses; evaluation plans; training effectiveness analyses (TEA), cost and training effectiveness analyses, and tests; and verification, validation, and accreditations (VV&A).
- Research Summaries include documents such as bibliographies, lessons learned, meta-analyses, and reviews.
- Methods are written guidance on how to conduct evaluations and for compiling measures of effectiveness or measures of performance.
- Policy documents contain DoD or Service guidance for conducting evaluations.

The literature review was built upon two earlier reviews, recommendations from SMEs, and a current review. The first of the earlier reviews was contained in the Training Effectiveness Catalogue System data base (Resource Consultants, Inc., 1992).[7] The other was conducted by DMDC during a survey of CEAT in the DoD (Simpson, 1995).[8] All of the documents in these two reviews were examined for relevance in this manual.

Subject-matter experts were requested via letter to identify case studies that illustrate good practice in the conduct of cost-effectiveness analysis, training effectiveness analysis, or cost analysis of training technologies and methods.[9] The request stated, in part:

> We solicit your help in identifying suitable case studies. These will typically be exemplary[10] R&D or test reports published in the last decade as technical reports or journal articles. The studies may focus on virtually *any type of military training technology or method in any context* (e.g., schoolhouse through unit training, training development or the conduct of training on an ongoing basis, classroom training or the use of training technology, use of small- through large-scale training simulations, individual or collective training). Further, they may focus on *any stage of development*, from initial conception through fielded system. Finally, each case study *must describe the methods it employs in sufficient detail that it can be applied by others.*

In addition, SMEs were asked to provide the rationale for their suggestions. The initial request was followed up with e-mail and phone calls. More than half the SMEs (or a colleague in the same organization) responded to the information request. SMEs recommended approximately three dozen evaluations for use as case studies.

Create Training and Cost-Effectiveness Data Base

The literature review was the basis for determining training and cost-effectiveness methods, procedures, and examples. A relational data base was created to organize information about these documents and their content. The resulting data base is called the Training and Cost Effectiveness File (TCEF). It was designed to serve two separate but related purposes: (1) analysis tool and (2) end product.

As an analysis tool, TCEF enables users to extract documents based on class; for example, evaluation method, example of evaluation, and DoD and Service training system evaluation policy. It permits a host of different types of data base searches. TCEF organizes a large body of information (500+ documents) and provides the means to make that information readily accessible to users.

[7] TECATs was developed in 1992 under DoD contract for the purpose of organizing then-current knowledge on training effectiveness evaluation. It contains information from approximately 400 reference documents.

[8] This review was based primarily on an electronic search of the Defense Technical Information Center (DTIC) data base to identify documents published between 1974 and 1994 relating to training effectiveness, cost analysis, cost and training effectiveness, cost-effectiveness, and various combinations of these and related terms. The review included approximately 1000 reference documents.

[9] Sixty SMEs were identified based on their contributions to the field of training effectiveness analysis in terms of publications and professional responsibilities. The author is indebted to Jesse Orlansky and Don Johnson for their help in compiling the list.

[10] What was meant by "exemplary" was left up to the SME. One SME did not like that this had been left open-ended: "[You should] specify characteristics that distinguish 'good' evaluations from the fake science that often passes for training effectiveness evaluations, rather than leaving it up to a panel of experts to nominate projects based on unknown, unspecified criteria." The vagueness was intentional. First, it would be difficult to define "exemplary" in terms that would be universally acceptable. There are many different evaluation methodologies (e.g., experiment, judgment, analytical, survey) and the standards of good practice in one do not necessarily apply in all others. Second, it was of interest to find out how the diverse audience of the survey—ranging from field testers to laboratory researchers—would agree or disagree on the meaning of the word. Finally, there was concern that imposing criteria of quality would artificially limit the number of responses to a few rare and exceptional case studies rather than a larger number of good but imperfect ones.

As an end product, TCEF is a historical training effectiveness data base that is available for user access.

TCEF indexes and summarizes key applied studies conducted by the Services and DoD to evaluate the training effectiveness and cost-effectiveness of various types of training (e.g., simulation, computer-based instruction, distance education, and military training programs). It can be used to identify representative studies, meta-analyses, and reviews; procedural guidance for conducting studies; and military requirements (e.g., directives, instructions) for conducting studies during training system development. Major data elements are type of document, citation, summary, abstract, training echelon, type, subtype, and content; and training evaluation method, submethod, level, and variables. TCEF will help users estimate the training effectiveness and cost-effectiveness of various types of training; identify procedural guidance for conducting evaluations; and identify examples of published evaluations.

User access to the TCEF data base became possible in February 1998. Contact the author for details.

Conduct Analyses

Several analyses were conducted to satisfy the manual's objectives. Most addressed questions relating to the why, who, what, where, how, and when of military training evaluations. Additional analyses were conducted to define and classify evaluation criteria. The analyses were conducted based on studies indexed in TCEF. These analyses are described in Chapter 2 (see Table 2-1).

Develop Evaluation Framework

This manual develops and describes a proposed training effectiveness evaluation framework for LSTS. For purposes of this manual, evaluation framework is defined as a set of evaluation principles and a description of evaluation events, their purpose, timing, and relevant dependent variables. The framework is intended to apply to any large-scale virtual, constructive, or advanced distributed simulation. The evaluation framework developed for this manual is intended to help the evaluator select the most suitable evaluation method based on the circumstances, provide procedural descriptions of the methods, and identify case studies. Case studies are examples of completed studies linked to each method that can be used as models to emulate. The framework was developed by integrating the concepts and information developed during the analyses. The framework may be thought of as a way to structure an evaluation based on underlying evaluation principles that enable evaluators to plan and time appropriate evaluation events.

The evaluation principles represent a philosophy toward evaluation. The principles reflect the hopes, standards, and reasonable expectations of the evaluator, given real-world constraints. They declare the evaluator's position on such matters as why evaluation is conducted, its intended effects on stakeholders, data quality expectations, and what data are deemed important.

Other Evaluation Considerations

DMDC work focused mainly on developing methodological guidance. DMDC identified areas in which policy changes would facilitate more effective LSTS evaluation. DMDC passed its findings to USD (P&R). Suggested policy changes following from DMDC analyses are not addressed in this manual.

Conduct SME Review

This manual was reviewed by SMEs from Service R&D laboratories, the operational testing community, Federally Funded Research and Development Centers, USD (P&R), and DMDC. The author tried to respond fully to comments received. Where conflicts remain unresolved, they are described in the text or footnotes.

Road Map: Where This Manual Addresses Each of Its Objectives

This manual contains the evaluation resources at the locations shown after each objective, below:

- Provide guidance to help analysts design meaningful training effectiveness evaluations: See Chapters 2 (Building an Evaluation Framework), 7 (Evaluation Criteria), and 8 (Evaluation Framework) for descriptions of the framework. See also Chapter 5 for discussion of evaluation problem areas.
- Describe procedures for alternative methods of conducting training effectiveness evaluations: See Chapter 3 (Evaluation Methods).
- Provide examples of training effectiveness evaluations that may be used as models to emulate: See examples of evaluation methods in Chapter 3; see Chapter 4 case studies of LSTS.

2 BUILDING AN EVALUATION FRAMEWORK

This chapter begins to develop and describe a training effectiveness evaluation framework for LSTS. For purposes of this manual, evaluation framework is defined as a set of evaluation principles and a description of evaluation events, their purpose, timing, and relevant dependent variables. The framework is intended to apply to any large-scale virtual, constructive, or advanced distributed simulation. The chapter begins by asking the most basic question about training evaluations; namely, why are they conducted? It then explores several related questions and begins to build an evaluation framework. The framework is described in detail in Chapter 8.

In thinking about training effectiveness evaluation, it is useful to start by asking basic questions; for example:

- What is the purpose of evaluation?
- What training treatment is being evaluated?
- What evaluation methods are used?

Questions such as these take on more specific meanings within an actual evaluation. If an evaluator starts by asking fundamentals, it is possible to avoid preconceptions and biases. To get and tell the whole story, the evaluator might start by following the newspaper editor's advice to the young reporter to get the why, who, what, where, when, and how. (However, get this not about a traffic accident or murder, but about military training evaluations.) These questions are posed in Table 2-1.

The right column offers examples and alternatives that might provide answers. For example, in response to the first question, the table offers four possibilities. In response to the second, it offers four; to the third, three; and so forth. Note that the second to last question (Evaluation criteria?) looks out of place in the editor's list. Here the evaluator must ask a question that the editor does not: What are the dependent variables? That is, what does the evaluator measure to judge training effectiveness? This chapter addresses each of the questions posed in Table 2-1, in turn.

Table 2-1. Some Key Evaluation Questions

QUESTION	QUESTION (EXPANDED)	EXAMPLES/ALTERNATIVES
Why?	What is purpose of evaluation?	• Satisfy milestone requirements • Identify design deficiencies • Resolve training problems • Predict training potential
Who?	Whose training is being evaluated?	• Individual • Team • Military organization • Joint force
What?	What is being evaluated?	• Virtual simulation • Constructive simulation • Advanced Distributed Simulation
Where?	Where is evaluation conducted?	• Laboratory • Field
How?	What evaluation methods are used?	• Experiment • Analysis • Judgment • Survey
Evaluation criteria?	What are the dependent variables?	• Reaction • Learning • Behavior/Processes • Results
When?	What is timing of evaluation events?	• Pre-development • During development • Post-development

Why Evaluate?

Large-scale training simulations are costly, complex, and difficult to evaluate. Because of their high cost, DoD regulations require them to undergo formal testing to see if they meet design objectives. The evaluation informs developers, decision-makers, and other stakeholders whether or not they deliver effective training, or less or more effective training, than an alternative, or provide equivalent training at reduced cost. Training evaluations are also conducted for different reasons at different points in time; for example, before development, to establish the need for a new or modified training system; during development, to refine the system; and post-development, to determine if training is relevant and useful on the job.

Training evaluations assess different ways to conduct training; for example, using alternative training methods, media, programs, and simulations. Evaluations are conducted for several different reasons during training development (prospective, developmental, milestone, post-development).[11] To illustrate, the reasons cited below are based on an analysis of 250 training evaluations in TCEF:

- Laboratory researchers and military trainers conduct TEAs to predict the training potential and effectiveness of new ways to train and to support the design of new training programs and systems.

[11] For purposes of definition, *prospective* refers to the pre-development phase, developmental to the phase during which the system is being developed, and *post-development* to the phase following development. *Milestone* is a significant contractually-required and scheduled developmental event; for example, demonstration of a functional capability.

- Training developers and training program managers conduct TEAs on ways to train that are undergoing development to satisfy military milestone[12] requirements, identify and correct system deficiencies, determine trainee and trainer preferences for certain features, improve designs, determine whether design standards are being met, and estimate training effectiveness.
- Trainers and training program managers conduct post-development TEAs on existing ways to train to resolve training problems, refine training, identify and correct training deficiencies, and determine overall training effectiveness.

There are many different specific reasons for conducting TEAs. Further, the reasons differ with stage of training development. The type and amount of evaluation data available depend on the developmental maturity of the training system. For example, at the prospective stage—before the system exists—evaluation is usually based on paper and pencil analyses and judgment data. At the various developmental stages (e.g., as when building a complex simulator over a period of years), the question is answered based on limited data at first and more data as the training system matures. Post-development, the question can be answered based on newly-generated and historical data.

Evaluators can conduct experiments in addition to conducting analyses and gathering judgment data. The more mature the training system, the more data are available and the more confidence in the evaluation.[13]

The response to evaluation varies with the audience and situation. For example, an evaluation may cause a researcher to modify a training concept to improve it or kill it. A program manager may decide whether to continue, modify, or terminate a training development. A schoolhouse trainer may recommend changes to an ongoing training program. There are many, many more possibilities.

Evaluations offer the opportunity to identify training system deficiencies and correct them. Evaluation conducted for this reason acknowledges that evaluation (1) is not an isolated event but a process, (2) is a technique to improve the system being evaluated, and (3) may or may not provide definitive results. In this sense, evaluation is a component of Total Quality Management (TQM). During TQM, data pertaining to a process are gathered and analyzed, the process is critiqued, and corrective actions are taken to improve the process. This goes on in an endless cycle.

[12] Training evaluation milestone requirements are set forth in DoD policy documents such as *DoD Directive 5000.1: Defense Acquisition*, which requires that evaluations be conducted during system development to assure that developmental systems demonstrate cost and operational [i.e., training] effectiveness. General DoD guidance is expanded in Service-specific policy statements such as *TRADOC Regulation 350-32: The TRADOC Training Effectiveness Analysis (TEA)* System, which suggests that TEAs may be conducted to determine training requirements, resolve training problems, and improve TEA study methodologies, as well as to assess training and cost-effectiveness of developmental systems.

[13] The downside to this proposition is that as the availability of training data increases, the potential to change a design decreases.

Whose Training Is Evaluated?

Individual Versus Collective Training

Most military training is directed at individuals to develop their individual skills. Large-scale training simulations are designed mainly to conduct collective training. Collective training is given to groups of individuals who work together and coordinate their activities. The size of the collective varies. The smallest collective is the team, usually with fewer than a dozen members. Its members might be the crew of an aircraft or military vehicle, or a command group consisting of senior officers who work together to wage a battle. A larger collective is a single-service military organization (e.g., battalion or brigade). Still larger collectives, consisting of joint or multi-service organizations, may participate in training with LSTS. The various members of these collectives are the "who" of the question in the title, above. They consist of the leaders, crew members, system operators, and others participating in training. An evaluation may involve more than one level of collective training.

Analyses were conducted on the 250 evaluations in TCEF to determine how they break down in terms of training echelon (individual, team, collective, joint), content area, and whether learning was classified as education or training.[14] Table 2-2 presents the results of these analyses.[15] The left-most column identifies echelon and also gives the percent of studies for each echelon.

Echelon

More TEAs were conducted for individual training (65%) than for team training (22%) or collective training (18%). None of the TEAs evaluated a joint training event, although such studies are conducted.[16] The table reveals the relative experience of the training evaluation community by content area. Strong areas for individual training are job skills, gunnery, flight, and education; overall, this echelon is well represented by evaluations and examples of representative studies are not hard to find. At the team echelon, there were many evaluations of military crews but few of command groups. At the collective echelon, combat is well represented. At both collective and joint echelon, there are no OOTW (Operations Other Than War) evaluations.

[14] Military *training* is formally defined as "instruction and applied exercises for the attainment and retention of skills, knowledge, and attitudes required to accomplish military tasks" (Department of Defense, 1990). Military personnel also undergo *education*, which is generally less applied than training and conducted to provide basic and advanced skills and knowledge to support professional development and advancement.

[15] Data are based on 250 evaluations in TCEF. Percentages were calculated based on frequency of occurrence of the training category divided by 250. As some evaluations involve more than one type of training, totals exceed 100%. Slightly fewer than 10% of these evaluations involved training at more than one echelon. In such cases, each echelon was counted separately.

[16] Joint training events are evaluated. However, the large scale and complexity of these events and the manner in which the data are analyzed and disseminated make their results far less accessible than traditional studies of individual, team, and collective training. The Joint category was included in TCEF as a place-holder and to provide for future growth.

Table 2-2. Training Content, Echelon, and Training Versus Education Taxonomy

ECHELON	CONTENT	EDUCATION VERSUS TRAINING	FREQ	PERCENT
I (Individual) (65% of cases)	job skills (e.g., air combat, combat engineer, maintenance, navigation, etc.)	training	62	25
	gunnery (e.g., Dragon, tank, TOW, Bradley, M16A1)	training	33	13
	education	education	27	11
	flight (e.g., fixed-wing, rotary-wing)	training	26	10
	combat leader (e.g., armor, aviation, naval)	training	6	2
	military system operation (e.g., assault bridge, ship, information system, sensor system, weapon system)	training	9	4
T (Team) (22% of cases)	military crew (e.g., aircrew, rifle squad, armor crew)	training	48	19
	command group (e.g., Army, Navy, USAF, Marines, Joint)	training	8	3
C (Collective) (18% of cases)	combat (e.g., air, ground, sea)	training	45	18
	OOTW	training	0	0
J (Joint) (0% of cases)	combat (e.g., air, ground, sea)	training	0	0
	OOTW	training	0	0

Content

The content column lists the categories and subcategories of training at each echelon derived from analyses. For example, the first category for Individual training is combat leader. This breaks down into subcategories for armor, aviation, and naval leader training. Other categories for individual training are flight, gunnery, military system operation, and job skills. Their related subcategories are self-evident.

Education Versus Training

Education is included at the individual echelon because most education occurs on an individual basis. There may be such a thing as team or collective education, but none of the TCEF studies revealed it. The categories and subcategories for the Team echelon are command group and military crew. A command group is a team of military leaders who work together to manage a military operation. A military crew is a team that mans a military vehicle, ship, or aircraft. The Collective and Joint echelons contain the same two subcategories: combat, and OOTW.

What Is Evaluated?

Taxonomy

Military training evaluations focus on different ways to train people. To illustrate, TCEF contains studies investigating the effects of computer-based instruction, self-paced instruction, simulator training, live gunnery training, and contracted aircrew training, to name a few. Analyses were conducted to develop a taxonomy to categorize the studies in terms of what they investigated. The taxonomy was built by selecting a sample of studies, developing a working set of training type categories, assigning studies by type, then expanding the sample, attempting to assign new studies to the types, modifying the types as necessary, expanding the sample further, and so on. Four different training types and subtypes provide a reasonable "fit" for all the studies in TCEF. The four training types are training medium, method, program, and simulation. The following definitions apply within this manual:[17]

- Simulation: Training tool that imitates one system or process with another.
- Program: Total system used to conduct training.
- Medium: Means to convey training without substantially altering its structure or content.
- Method: Particular way, technique, or process used to train.[18]

Table 2-3 presents the results of these analyses for the 250 TCEF evaluations. The left-most column identifies training type and also gives the percent of studies for each type. The percent of TEAs is largest for simulation (54%), smaller for program (22%) and medium (21%), and smallest for method (6%).[19] Table 2-3 also shows the training subtypes for each type and the percent of cases each represents.

Simulations

More than half of the TCEF evaluations deal with simulations. Most are of virtual simulations. Gunnery, crew, and flight simulations are also well represented. The widespread interest in virtual simulations such as SIMNET and CCTT is reflected in the number of evaluation studies. Few evaluation studies on constructive simulations have been published.

[17] Various definitions of these terms exist within the educational literature and they are somewhat contradictory and overlapping. The definitions used here are based on those in Merriam-Webster (1986): medium ("means of effecting or conveying something"), method ("way, technique, or process of or for doing something"), and program ("plan or system under which action may be taken toward a goal"), simulation ("imitative representation of the functioning of one system or process by means of the functioning of another").

[18] Glaser (1976) defines method as "conditions which can be implemented to foster the acquisition of competence."

[19] Percentages are calculated based on frequency of occurrence of the training subtype divided by 250. As some evaluations involve more than one subtype, totals exceed 100%.

Table 2-3. Training Type and Subtype Taxonomy

TRAINING TYPE	TRAINING SUBTYPE	FREQ	PERCENT
simulation (54% of cases)	virtual	51	20
	gunnery	30	12
	crew	16	6
	flight	16	6
	live	7	3
	operator	6	2
	constructive	8	3
	maintenance	2	1
program (22% of cases)	N/A	55	22
medium (21% of cases)	ITV (instructional TV)	16	6
	ICAI (intelligent computer-aided instruction)	10	4
	CBI (computer-based instruction)	12	5
	IVD (interactive video disk)	9	4
	other	6	2
method (6% of cases)	various	14	6

Exactly what gets evaluated at a particular point in system development depends on what is available. During early development, evaluators must work with mature subsystems, mockups, "breadboard" simulations,[20] published specifications, or other representations that can be analyzed and manipulated. The total system cannot be evaluated until fully developed; this might not be possible until several years after development begins.

Programs

Roughly one-fourth of the evaluations deal with training programs. In most cases, these programs have been operational for years. The studies were usually conducted to validate program effectiveness.

Media

In the medium training type, the most common subtypes are CBI (Computer-Based Instruction), ICAI[21] (Intelligent Computer-Aided Instruction), ITV (Instructional TV), and IVD (Interactive Video Disk). There is also an Other subtype, for training media that do not fit into the four other subtypes. Some media that did not fit are embedded training, exportable training materials, Internet, and training aid. The number of medium subtypes shown in Table 2-3 is small because most of the studies in TCEF were conducted in the last decade where these four subtypes were the focus of research interest.[22]

Methods

The method training type is the smallest in this sample. This is due to the applied nature of military training research and intentional

[20] A breadboard simulation might be thought of as a quick prototype simulation of the simulation.

[21] Some proponents of ICAI refer to their systems as Intelligent Tutoring Systems (ITS) (Shute, 1991; Shute and Psotka, 1994). Fletcher (1988) makes the argument that ICAI must be able to represent (1) the knowledge domain, (2) student's state of knowledge, and (3) an expert tutor. Conventional CBI does not have to meet all of these requirements. If ICAI can represent an expert tutor, then it is by the definition used in this manual a training *method* versus a *medium*. The author takes the position that while some ICAI may meet the expert tutor requirement, this is by no means the general case. Hence, ICAI has been assigned to the medium training type.

[22] The possible media include essentially any technique that can be used to transmit and display training information. To give a sense of some of the possibilities, a 1982 evaluation of the media selection process (Kribs and Mark) identified the following media alternatives: lecture, programmed text, linear text, workbook, programmed filmstrip, slide with sound, random-access slide, videotape cassette, videodisc, computer-controlled instructional TV, PLATO, General Electric Training System, and microfiche.

selection of applied versus theoretical studies for inclusion in TCEF. The method category would be larger and more diverse in a sample of academic (i.e., theoretically-oriented) research studies. Moreover, method is broadly enough defined for purposes of this manual that it could include instructional strategies. Evaluators can gain a sense of the potential breadth of this topic by considering the training methods and strategies covered in recent academic reviews.[23] The DoD and Services sponsor a larger amount of this research than is contained in TCEF.

Where To Evaluate?

The main question here is whether to evaluate in the laboratory or in the field. The "field" is the normal work setting of military personnel and their equipment. This might be on board ship, within a military unit, in a military classroom, or other place that troops operate. A "laboratory" is an artificial setting where evaluators exercise a high degree of control over extraneous variables. The distinction between laboratory and field is not as much one of geography as of control. Evaluators exercise more control over "laboratory" studies and less over "field" studies, regardless of the setting. Most military training evaluations are field evaluations.[24] Field evaluations pose special problems to the evaluator, although they are acknowledged to possess greater external validity. Military training evaluations usually take place in the field. This topic is discussed in greater detail in Chapter 5.

How To Evaluate?

Perspectives

To ask how to evaluate is to ask what evaluation methods to use. There are many of these, and making the choice is not always straightforward. Evaluation means different things to different people, depending upon their background and experience. To illustrate, here are some different evaluation communities and the methods commonly associated with each:

- Academic or Service laboratory: Conduct a laboratory experiment.
- Operational Test & Evaluation: Conduct a field test to see if system meets design standard.
- Military trainer: Compare student test scores before and after training.
- Military decision-maker: Get judgments of military end users and other experts.
- Operations research: Determine predictions of mathematical models.

[23] For example, tutoring (Bloom, 1984); mastery learning (Kulik, C.-L.C., Kulik, J.A., and Bangert-Drowns, R.L., 1990); programmed instruction (Kulik, C.C., Schwalb, B.J., and Kulik, J.A., 1982); accelerated instruction (Kulik, J.A and Kulik, C.-L.C., 1984); Keller's personalized system of instruction (Kulik, J.A., Kulik, C.C., and Cohen, P.A., 1979); effects of advance organizers (Luiten, J., Ames, W., and Ackerson, G., 1980); classroom reinforcement (Lysakowski, R.S., and Walberg, H.J., 1981); instructional effects of cues, participation, and corrective feedback (Lysakowski, R.S., and Walberg, H.J., 1982); effects of homework on learning (Paschal, R., Weinstein, T., and Walberg, H.J., 1984); teacher questioning behavior (Redfield, D.L. and Rousseau, E.W., 1981); cooperative learning (Slavin, R.E., 1980); use of instructional systems (Willett, J.B., Yamashita, J.J., and Anderson, R.D., 1983). Some instructional strategies that have been the subject of academic research are: self interrogation, note-taking, imagery, chunking, verbalization, guided writing, small group brainstorming, networking, peer learning, paraphrasing, question-answering, visual imagery, pretraining, mnemonic techniques (method of loci, absurd pictures, narratives, rhymes, acronyms, acrostics, numerical acrostics, graphic illustrations, spontaneous associations), self-monitoring, SQ3R method for studying (survey, query, read, reread, recall). In general, topics such as these tend to be covered in theoretically-oriented (i.e., academic) research and are seldom covered in military training studies, which tend to be more applied.

[24] Virtually all of the evaluations in TCEF are field evaluations. This is due to the generally applied nature of military training research and to the intentional selection of applied versus theoretical studies for inclusion in TCEF. A half-dozen or so TCEF evaluations are on the borderline between laboratory and field but were classified as field evaluations.

No community's take on evaluation is right or wrong. It is important to recognize that different schools of thought exist, however. Not doing this can cause communication problems when different evaluation communities interact. More serious, however, is that preconceptions tend to make evaluators myopic; that is, unable to see another community's point of view about how to conduct an evaluation. No single methodology is suitable for all training effectiveness evaluations. Methods vary in terms of applicability at different stages of system development, amount and quality of the data they provide, cost, and other factors. One of the functions of an evaluation framework is to help the evaluator decide what method(s) are most appropriate at different stages of training system development.

Levels of Evaluation

Different classes of training evaluation methods differ in their procedures and the levels of evaluation data they generate. Jeantheau (1971) distinguishes among these four levels of evaluation:

- Qualitative
- Non-comparative
- Comparative
- Transfer

Move down the list from top to bottom and the data gain authority. Qualitative evaluation is based on subjective estimates that do not assign quantitative value. For example, an evaluator might rank a training system attribute as "good" but be unable to say how good in any absolute sense.

Non-comparative evaluation assigns value based on a set of standards. This is often done during training system development. Quantitative value can be assigned. An example would be to conduct training on a simulator and to rate its effectiveness based on the percent of training tasks students perform to standard.

Comparative evaluation assigns value to two or more competing training alternatives. Quantitative value can be assigned. At the end, the winner can be picked based on the values obtained.

Transfer evaluation assigns value based on performance in a new situation. An example is transfer from a flight simulator to performance piloting an aircraft. If two alternatives are being compared, the winner is the one with the greatest percent of transfer.[25]

[25] Here is a more down-to-earth example of the practical significance of these levels. Consider the aspiring athlete who is considering the purchase of a piece of exercise equipment to aid preparation for an important athletic event. In the local athletic equipment superstore, he or she examines the alternative devices. In making the decision on which one to buy, likely the first choice will be based on purely subjective reasons; for example, the cachet associated with a brand name. At second glance, the athlete may check a list of features. At third, he or she may compare device A with device B with device C, and so forth. Finally, the athlete may ask what evidence, if any, demonstrates that use of a particular device influences performance in the athletic event. At each succeeding level, the question asked more clearly addresses the true value of the device to the athlete.

Taxonomy

What evaluation methods may be used to conduct a TEA? The first step in answering this question is to develop a descriptive taxonomy of alternative methods that have been used in military training evaluation. Evaluations in TCEF tend to use one of four main methods: experiment, judgment, analysis, or survey.[26] [27] In general terms, here is how the methods are applied:

- Experiments determine effectiveness based on observational[28] data.
- Judgment-based evaluations determine effectiveness based on human judgments.
- Analytical evaluations determine effectiveness based on common analytical techniques and using common analytical strategies.
- Surveys gather data from a sample of a knowledgeable target population and determine effectiveness based on analysis of the collected data.

Each of the methods can be performed in several different ways, comprising a set of submethods. The submethods of Experiment are defined mainly based on distinctions made in Campbell and Stanley (1966).[29] The submethods of Judgment are based on respondent category; that is, the group whose judgments are considered (Users, SMEs, or Analysts).[30] The submethods of Analysis are based on differences in the objectives of analysis (Evaluate, Compare, Optimize).[31]

The submethods for these methods were developed iteratively based on analysis of TCEF evaluations. The distinctions are based on differences in use. The Survey method has no submethods. A larger sample of surveys would permit submethods to be defined, but no useful distinctions could be made based on the 14 surveys in TCEF. The submethods vary in terms of cost, difficulty, and the authority with which they support conclusions. This subject is discussed in Chapter 3.

[26] This terminology is used to facilitate discussion of the different "methods." Be aware that the language simplifies. First, experiment and analysis reasonably fit the dictionary definition of method (e.g., a systematic procedure). Strictly speaking, *judgment* is a type of data and *survey* is a means of data collection. Judgment and survey are referred to as methods in this document because they tend to be used in certain predictable ways that comprise systematic procedures that, in fact, constitute methods. However, these definitions do not necessarily generalize outside the pages of this manual.

[27] Caveats: (1) taxonomy is based on a historical record of 250 actual evaluations; if a smaller or larger sample were used, or if the sample were extended to evaluations published in the academic literature, the taxonomy might look different; (2) if an evaluation theorist attempted to construct such a taxonomy from, say, academic literature on evaluation, it would also look different; (3) the taxonomy is intended to provide a simple framework for discussion. What can be said about this taxonomy is that it reasonably represents common practice in the conduct of military training evaluations, for better or worse.

[28] Experiments may also use judgment data; for example, gathering the judgments of two different groups participating in an experiment.

[29] Campbell and Stanley do not define *test* or *transfer*. For our purposes, these definitions apply: Test is a single-group experiment in which success is judged against a *predefined standard. Transfer* is an experiment that attempts to measure the effects of learning in one situation (e.g., using a flight simulator) to performance in another (e.g., flying an aircraft).

[30] Another way to break down judgment is by the method used to collect the data; for example, questionnaire, survey, interview, user comments recorded by observers, critical incident reports (Bessemer, 1998, 13 August). The author used respondent category because it implicitly states something about the authority of the judgment.

Table 2-4 breaks down the frequency of use of methods and submethods in TCEF. These data show the frequency with which each method and submethod is the primary method used in an evaluation. The method most commonly used is experiment (65% of cases). Judgment (13%), Analysis (17%), and Survey (6%) are used in far fewer cases. Different methods are sometimes used in combination, although one of the methods is usually primary. Pairings of experiment, analysis, and survey are rare, but judgment data are often used with other methods. Judgment data are often gathered in combination with experiment.[32]

Why do these numbers differ? Some possible reasons:

- Acquisition regulations encourage experiments.
- Among most evaluators and military decision-makers, experiments have greater face validity than other methods.
- Analysis- and Judgment-based evaluations are less difficult and costly than experiments and so tend to be used when experiments are not possible.

Table 2-4. Frequency of Usage of Common Evaluation Methods and Submethods

METHOD	SUBMETHODS	FREQ	PERCENT
Experiment (65% of cases)	True experiment	72	29
	Transfer	22	9
	Pre-experiment	24	10
	Test	16	6
	Quasi-experiment	12	5
	Ex post facto	15	6
Judgment (13% of cases)	Users	15	6
	SMEs	12	5
	Analysts	5	2
Analysis (17% of cases)	Evaluate	26	10
	Compare	11	4
	Optimize	6	2
Survey (6% of cases)		14	6

Timing

To use experiment, a training system must exist and be functional in some form.[33] Judgment can be used before a system exists (e.g., to estimate training potential of a hypothetical design or the perceived need for a system), but usually requires a functional system. On the other hand, analysis can be performed without an existing training system. Analysis tends to be used in two main cases:

- The system is not developed enough to conduct an experiment or gather judgment data.
- Evaluation resources are limited.

[31] As with judgment data, submethods could have been defined based on other methodological characteristics; for example, the use of analytical strategies such as modeling, analogy, extrapolation, task list analysis, historical data. See Chapter 3 for further discussion of this subject.

[32] Table 2-4 shows the frequency with which Judgment was the primary method used (32 cases). TCEF includes a total of 76 evaluations in which Judgment was used, indicating 44 cases of use in combination with other methods. If one counts all the uses of Judgment, it represents nearly one-third of the evaluations.

[33] The system does not necessarily have to be actual, complete, or final. In some cases, it may be possible to use a mockup or simulation to represent the system. Enough of the system must be represented to conduct a meaningful experiment.

Experiments are conducted with real versus hypothetical things. This is usually true of surveys. An exception to this rule is that a survey might be conducted to determine the need for a new training program.

Based on TCEF, analytical evaluations are about twice as likely to be performed on hypothetical as on existing ways to train. For judgment-based evaluations, the opposite is true. This sample suggests that hypothetical ways to train are more commonly evaluated analytically than based on judgment.

Levels, Revisited

The different evaluation methods and submethods are usually used in ways that yield different combinations of Jeantheau's four levels. Table 2-5 illustrates the levels of data commonly associated with the different evaluation methods. In principle, all boxes could be checked because it is possible to obtain data at all four levels using all methods. In practice, however, the methods tend to be used more narrowly. Based on the evaluations in TCEF, it can be said that experiments are most often used to provide comparative data and second most likely to provide non-comparative or transfer data. In principle, they can be used to provide qualitative data. However, this did not occur in any of the TCEF evaluations. A Survey may be thought of a type of experiment that yields qualitative data.

Judgment was used to obtain qualitative, non-comparative, and comparative data. It was about twice as likely to be used qualitatively or non-comparatively as comparatively.[34]

Judgment was never used to obtain transfer data, although this is possible (see Chapter 3).

Analysis was used non-comparatively and comparatively but never qualitatively or to estimate transfer. Analysis to obtain qualitative data is possible. It is also possible to use analysis to estimate transfer. Survey was always used qualitatively or non-comparatively but never comparatively or to estimate transfer; both are possible but improbable.[35]

These methods tend to be used in certain predictable ways to gather different levels of data. Because the levels differ with method, each method has inherent strengths and weaknesses when compared with other methods. Chapter 3 discusses evaluation methods in more detail.

[34] Note that this is for evaluations in which judgment was the primary evaluation method used. When used with other methods—for example, experiments—it was more likely to be used comparatively.

[35] None of the surveys in TCEF was used to obtain comparative data. This is possible, for example, by asking a group of respondents to estimate the relative training effectiveness of two different ways to train. Within public opinion polling, comparative surveys are common; for example, a political survey that asks respondents who of a group of candidates they intend to vote for.

Table 2-5. Levels of Data Commonly Associated with Evaluation Methods

METHOD	TYPE OF DATA			
	Qualitative	Non-Comparative	Comparative	Transfer
Experiment		√	√	√
Judgment	√	√	√	
Analysis		√	√	
Survey	√	√		

What Are Evaluation Criteria?

Evaluation criteria are the measures collected during an evaluation whose values are used to decide the outcome of the evaluation. Dependent variables in experimental research are one type of evaluation criteria.

Reactions

The simplest variable to measure is reactions of participants to a training experience. This is done with a post-training questionnaire, interview, or videotaped group discussion such as an AAR (After-Action Review).

Combat Performance

The operational testing community emphasizes the use of measures of *combat* performance such as engagement or battle outcomes. Many of the evaluations this community performs are of weapon systems and the concern with combat outcomes is obvious. There are analogous variables for training systems. First, the evaluator could measure trainee performance during the simulation in relation to combat objectives, For example, did the simulated tank company defeat the simulated enemy; or, did the senior commanders participating in a war game win the war? Second, the evaluator might want to measure transfer of training from the simulation to the real world. This could be done at a number of removes from the simulation. One way would be to measure the impact of simulator training on performance in live simulation; for example, performance of Army units at the National Training Center (NTC), or Navy forces in fleet exercises, and so forth. Another way—more difficult, but more persuasive yet—would be to determine the impact of simulator training performance in actual combat.

Student Learning

It is common to evaluate student performance in the schoolhouse based on test scores. Standardized tests can be used to evaluate training effectiveness. Students in an effective schoolhouse will achieve passing test scores. Can such test scores be used to evaluate LSTS? In theory, yes, but practically, no. First, the schoolhouse focuses on individual training whereas an LSTS focuses on collective training. Second, there are no collective tests or collective test scores in a simulation. There *is*, however, collective performance. Improvement in collective performance demonstrates learning. Hence, LSTS could be evaluated based on collective learning.

Collective performance is visible in process rather than singular events. Measurement of this process is difficult and special techniques are required (see Chapter 6). Moreover, a simulation may have several different collectives functioning simultaneously; for example, vehicle crews, command groups, companies, battalions, and so forth.

Collective Task Performance

Training—of individuals or collectives—is built upon tasks. Large-scale training simulations provide training on collective tasks. At the Joint level, the Universal Joint Task List (UJTL) describes the tasks that are to be performed by a joint military force, the conditions under which the tasks are performed, and standards of performance. Comparable Service-specific task lists define the relevant collective tasks at the Service level. These task lists define what tasks the Services and Joint forces are expected to be able to perform. They are the logical tasks to use when building scenarios to evaluate LSTS. Because they are hierarchical, they define tasks at more than one collective level. Hence, they suggest sets of dependent variables at each level.

Chapter 7 discusses evaluation criteria in greater detail.

When To Evaluate?

Evaluations are conducted for different reasons at different stages of system development. For example, pre-development, evaluations are conducted to determine whether a prototype design can train on certain tasks. During development, evaluation is conducted to identify and correct system deficiencies. Post-development, evaluation is conducted to determine whether training influences unit readiness.

DoD Directives and Regulations

The DoD acquisition rules acknowledge that system development is complex, lengthy, and expensive. To help system developers, acquisition directives and regulations (e.g., Department of Defense, 1996a,b) promote an orderly succession of developmental phases:

- Phase 0: Concept exploration
- Phase I: Program definition and risk reduction
- Phase II: Engineering and manufacturing development low rate initial production
- Phase III: Production, fielding/deployment, and operational support

Developmental Phases

Phase 0 (Concept Exploration) typically consists of competitive, parallel short-term concept studies conducted to define and evaluate the feasibility of alternative concepts. During Phase I (Program Definition and Risk Reduction), the program explores one or more concepts, design approaches, and/or parallel technologies to determine their advantages and disadvantages. Prototyping, demonstrations, and early operational assessments may be conducted. The primary objectives of Phase II (Engineering and Manufacturing Development) are to refine the design, demonstrate system capabilities through testing, and work out the manufacturing process. Low Rate Initial Production then produces a quantity necessary for operational tests and to establish a production baseline. The objectives of Phase III (Production, Fielding/Deployment, and Operational Support) are to achieve an operational capability that satisfies mission needs. Deficiencies encountered in testing are corrected and a support program is established. Follow-on operational testing occurs to assess system performance, and deficiencies are corrected.

Corresponding milestone decision points occur prior to each phase:

- Phase 0: Approval to conduct concept studies
- Phase I: Approval to begin a new acquisition program
- Phase II: Approval to enter engineering and manufacturing development
- Phase III: Production of fielding/deployment approval

Milestone Decision Points

Milestone decision points are established early in the program. At each milestone a program review is conducted to determine whether or not the program is progressing satisfactorily. If the outcome of a milestone test is unsatisfactory, the decision may be made to terminate development. Milestone 0 (Approval to Conduct Concept Studies) reviews the mission needs statement (MNS), identifies possible materiel alternatives, and authorizes concept studies, if they are deemed necessary. (A favorable Milestone 0 decision does not guarantee that a new acquisition program has been initiated.)

The purpose of the Milestone I (Approval To Begin a New Acquisition Program) decision point is to determine if the results of Phase 0 warrant establishing a new acquisition program and to approve entry into Phase I.

The purpose of the Milestone II (Approval To Enter Engineering and Manufacturing Development) decision point is to determine if the results of Phase I warrant continuation of the program into Phase II.

The purpose of the Milestone III (Production or Fielding/Deployment Approval) decision point is to authorize entrance into production or into deployment.

Development As Process Versus Event

Evaluations are often thought of as one-shot events that answer a question at a particular point in time. This may make sense when evaluating simple things that already exist, such as an inexpensive training method or medium. It does not make sense when evaluating complex and expensive LSTS that undergo years of development before becoming operational. Here, evaluation may occur as a series of several relatively small evaluation events, culminating periodically in larger milestone events, and eventually in a Phase III evaluation. Given that evaluation cannot be completed in a single stroke, the question is how to structure a series of events that will support the development and fielding of the most-training effective simulation. These events and their timing are discussed in greater detail in Chapter 8.

3 EVALUATION METHODS

This chapter describes the methods commonly used in military training effectiveness evaluations. The methods were mentioned in Chapter 2 (Table 2-4) but not described in detail. The present chapter should familiarize the reader with how training evaluation has been conducted as well as case studies.[36] The discussion is organized based on the four-method taxonomy (experiment, judgment, analysis, and survey) in Chapter 2.

The descriptions are based on the 250 evaluations in TCEF and are illustrated with many concrete examples. The evaluations in TCEF are a representative sample. Many but not all of these studies were recommended by SMEs. Some studies contain methodological flaws. Though imperfect, they represent the real world. Chapter 5 (Evaluation Problem Areas) identifies common flaws in evaluations and may help the reader judge the examples in this chapter.

About one-fourth of the evaluations in TCEF relate directly to LSTS. Most do not. However, most of the methods and examples in this chapter are still relevant when evaluating LSTS.

This chapter provides descriptive statistics on the relative use of the different evaluation methods. Please do not equate relative use with value or as a prescription for future use. These numbers simply tell how evaluations have been conducted in the past.

The chapter discusses each of the four classes of methods, and their submethods, in turn.

Experiments

For purposes of discussion, define experiment as an activity during which observational data are gathered. Observational data are usually objective, but may be subjective.[37] A definition this general is required to include the wide range of studies that evaluators call "experiments." This method breaks down further into submethods; that is, the method's family members.

Three submethods in the taxonomy match categories defined by Campbell and Stanley (1966):[38] true experiment, pre-experiment, and quasi-experiment. The remaining submethods are defined separately for purposes of this manual.[39]

[36] References provides complete citations of every publication cited in this manual. One of the selection criteria for including a document in TCEF was its ready availability. Technical reports may be obtained from DTIC and journal articles from technical reference libraries or via interlibrary loan.

[37] It depends upon what is used as the dependent variable. For example, in an experiment comparing two different ways to train tank gunners, the comparison might be made based on objective measures such as gunnery scores, subjective measures such as gunner ratings of the quality of the gunnery system, or both.

[38] If the reader is unfamiliar with these authors, it would help to review their key writings. Check References for Campbell and Stanley (1966), Cook and Campbell (1976), and Cook and Campbell (1979). The 1966 work (an 84-page book) was first published in 1963 in N.L. Gage (Ed.) *Handbook of Research on Teaching*. From the earlier to the later publications, each of these is a successively more elaborate treatment of ideas presented in the earlier works. This manual cites primarily the 1966 work, based on its convenience and brevity.

[39] There are some overlaps in the categories used in the taxonomy. The simplifications are made to facilitate discussion of a complex subject at a general level.

True Experiment

Campbell and Stanley (1966) describe three "true experimental designs"; these designs compensate for confounds likely to reduce internal and external validity (Table 3-1).[40] The designs share in common (1) use of a control group and (2) random assignment of subjects. In all cases, the dependent variable is measured with a posttest, although Design 6 does not use a pretest. Designs 4 and 6 use two groups and Design 5 uses four groups. Campbell and Stanley note that researchers may be reluctant to give up the pretest but state that "...within the limits of confidence stated by the tests of significance, randomization can suffice without the pretest" (p. 25). In other words, if subjects are randomly assigned, the pretest is optional. The templates shown in Table 3-1 can be extended to include more complex designs (e.g., factorials), provided the extensions incorporate comparable control groups, random assignment, and testing.

Table 3-1. Three True Experimental Designs (from Campbell & Stanley, 1966)

DESIGN	Campbell & Stanley	DESCRIPTION	CONTROL GROUP?	RANDOM ASSIGNMENT?	PRETEST-POSTTEST?	FREQ	PER-CENT
4	Pretest-posttest control group	R O X O R O O	yes	yes	yes	22	31
5	Solomon 4-group	R O X O R O O R X O R O	yes	yes	yes	0	0
6	Posttest-only control group	R X O R O	yes	yes	no	50	69

Legend: R (randomization of subjects), O (measurement of dependent variable), X (experimental treatment)

Seventy-two studies in TCEF were classified as true experiments based on methodological descriptions of control groups, testing, and random assignment of subjects. Random assignment was clearly specified in only a fraction of the studies. Hence, it was unclear whether random assignment had occurred or whether the designs were compromised versions of true experiments. Because it is often difficult to randomly assign subjects in field studies, it is reasonable to assume that random assignment had not occurred in many of these studies.

The relative use of Designs 4, 5, and 6 is shown in Table 3-1. The frequency of use of these designs appears to be inversely related to complexity. Design 5, which uses four groups and requires both pre- and posttesting, was never used. Design 4, which requires both pre- and posttesting, was used more than twice as often as Design 6, which requires only posttesting. The most frequently

[40] Internal and external validity are discussed in Chapter 5. For purposes of discussion here, the terms can be interpreted as follows: *internal validity* (can you predict the outcome based on the treatment?) and *external validity* (does the outcome generalize to other populations, settings, and variables?).

used design was Design 6, which is the most likely to lack randomization.

Examples of studies using these designs:

- Design 4: Brown, Pishel, and Southard (1988). A 2-group experiment. Eight platoons were pretested, then four each participated in SIMNET and field training, then all participated in ARTEPs (Army Test and Evaluation Program) as posttest.
- Design 4 with additional judgment data: Wetzel-Smith, Ellis, Reynolds, and Wulfeck (1995). A 3-group between-subjects experiment. Two different IMAT (Interactive Multisensor Analysis Training) groups and a control group (conventional training) were pre-tested, underwent training, and were posttested. Judgment data were also obtained.
- Design 5: None.
- Design 6: Greene and Haynes (1988). A 2-group experiment. Groups of TOW (Tube-Launched, Optically Tracked, Wire-Guided Missile) gunners were trained with two different types of gunnery simulators and then tested on TOW live firing.
- Design 6 with additional judgment data: Simpson, Wetzel, and Pugh (1995). A 3-group experiment. Students were pretested, participated in training, SMEs rated their performance, and at the end of the course were posttested and instructor and student attitude and judgment data were gathered.

Other examples of studies using these designs are listed in Reference List A-1 in Appendix A.

Pre-Experiment

Campbell and Stanley (1966) describe three "pre-experimental designs." These designs lack controls necessary for internal and external validity. The designs are summarized in Table 3-2. None uses randomization. Design 1 uses no pre- or posttesting or control group. Design 2 uses pre- and posttesting but no control group. Design 3 uses a control group but no pre- or posttesting. Studies using these designs are flawed but not uncommon in the published literature.[41]

[41] Campbell and Stanley critique these designs severely. Studies using Design 1 "have such a total absence of control as to be of almost no scientific value [because] securing scientific evidence involves making at least one comparison" (p. 6). For Design 2, they describe the factors that may intervene between pre- and posttest to produce confounded effects: *history* (occurrence of events other than X), *maturation* (systematic variation of subject psychological/biological processes unrelated to X), *testing* (effect of pretest), *instrumentation* (possible change in measurement instrument between pre- and posttest), and *statistical regression toward the mean.* Studies using Design 3 are flawed because they do not prove that the two groups would have been equivalent without X.

Table 3-2. Three Pre-Experimental Designs (from Campbell & Stanley, 1966)

DESIGN	Campbell & Stanley	DESCRIPTION	CONTROL GROUP?	RANDOM ASSIGNMENT?	PRETEST-POSTTEST?	FREQ	PER-CENT
1	1-shot case study	X O	no	no	no	6	25
2	1-group pretest-posttest	O X O	no	no	yes	17	71
3	Static group comparison	X O O	yes	no	no	1	4

Legend: O (measurement of dependent variable), X (experimental treatment)

Twenty-four studies in TCEF were classified as pre-experiments based on their methodological descriptions. Of the 24 studies, the frequency and percent of studies of Designs 1, 2, and 3 are shown in Table 3-2. Design 1 was used six times.[42] Design 3 was used once. Design 2 was used relatively frequently, even compared to true and transfer experiments (see Table 2-4). Evidently, many training evaluators believe that useful information can be obtained from Design 2. A common form of Design 2 is the in-device learning experiment, a relatively low-cost way for system developers to test whether a new training device is effective for training. This design is also commonly used to evaluate innovations in schoolhouse training. Studies using Design 2 use fewer subjects and are less complex and costly than true experiments. Evaluators who conduct such studies implicitly discount the confounding factors Campbell and Stanley have identified.

Examples of studies using these designs:

- Design 1: Harris (1996). Description of BFTT (Battle Force Tactical Trainer) developmental test. Training data were collected using team training assessment methodology (developed by Naval Air Warfare Center)—driven by scenarios and based on ratings of individual and team performance on process and product measures and followed by comprehensive debriefs.
- Design 2: Lampton (1989). Platoon leaders participated in one tactical exercise with Simulation in Combined Arms Training (SIMCAT) (pretest), were trained with SIMCAT, then participated in a third exercise (posttest). Performance was evaluated using AMTEP (ARTEP Mission Training Plan) standards.
- Design 2 with additional judgment data: Harman, Bell, and Laughy (1989). Pre- and post-test learning experiment. Twenty-seven combat engineers took pretest on mathematics skills, used tutor, then took posttest and attitude questionnaire.

[42] Design 1 is procedurally equivalent to the *Test* submethod, although its objectives and underlying assumptions differ. Refer to discussion of *Test*, later in this chapter.

- Design 3: Pleban, Brown, and Martin (1997). Sixteen subjects were assigned to one of two groups. Experimental group received the CBI version of the Principles of War module and an end of module quiz. Subjects assigned to the control condition received only the end of module quiz.

Other examples of studies using these designs are listed in Reference List A-2 in Appendix A.

Quasi-Experiment

Campbell and Stanley (1966) describe 10 different "quasi-experimental designs" (Table 3-3). These usually lack the controls of "true" experiments. The evaluator has little or no control over X and may not be able to use control groups, randomly assign subjects, or conduct pre- and posttests. These shortcomings reflect the natural circumstances in which quasi-experiments usually occur. Most of these designs reflect a set of events that presents a data collection opportunity. The quasi-experiment is not so much designed as it is recognized. Quasi-experiments can provide useful data if the evaluator can find suitable ways to compensate for their limitations.

Twelve studies in TCEF were classified as quasi-experiments based on their methodological descriptions. Of the 12 studies, one was Design 7, time series experiment, and the remaining 11 were Design 8, equivalent time samples design.[43] No control groups were used. In each of the Design 8 experiments, the performance of subjects was measured at intervals during their interaction with a training device or simulator. Learning curves for performance were generated as a function of exposure to the device. Most of these studies were conducted under conditions in which a control group was impractical. Use of a quasi-experiment made it possible to gain useful information on performance growth within the simulator.

Why were none of the other quasi-experimental designs used? One possible reason is that the designs are too unusual for military training research. Another possibility is that military training evaluators are unfamiliar with the designs. Complexity does not appear to be a problem with Designs 7 (time-series experiment), 10 (nonequivalent control group), or 10 (regression discontinuity). Design 10 is similar to ex post facto designs, discussed later in this chapter.

[43] These numbers are slightly misleading because six of the studies report on two of the same evaluations from the perspectives of different author-participants. If one counts evaluations instead of published evaluation studies, the number using Design 8 drops from 11 to 7.

Table 3-3. Ten Quasi-Experimental Designs (from Campbell & Stanley, 1966)

DESIGN	Campbell & Stanley	DESCRIPTION	CONTROL GROUP?	RANDOM ASSIGNMENT?	PRE- & POST TEST?	FREQ	PER-CENT
7	Time-series experiment	O O O O X O O O O	no	no	yes	1	8
8	Equivalent time samples design	X_1O X_0O X_1O X_0O	no	no	yes	11	92
9	Equivalent materials samples design	$M_a X_1$0 $M_b X_0$O $M_c X_1$O $M_d X_0$O, etc.	no	no	yes	0	0
10	Nonequivalent control group design (common in naturally assembled collectives; pretest deals with nonequivalence)	O X O O O	yes	no	yes	0	0
11	Counterbalanced designs (all participants experience all treatments)	X_1O X_2O X_3O X_4O X_2O X_4O X_1O X_3O X_3O X_1O X_4O X_2O X_4O X_3O X_2O X_1O	yes	no	yes	0	0
12	Separate sample pretest-posttest design (reasonable when applied to large populations)	R O (X) R X O (R=randomly equivalent subgroups)	yes	yes	yes	0	0
13	Separate sample pretest-posttest control group design	R O (X) R X O R O R O	yes	yes	yes	0	0
14	Multiple time series	O O OXO O O O O O O O O	yes	no	yes	0	0
15	Institutional cycle design	Class A X O_1 Class B_1 RO_2 X O_3 Class B_2 R X O_4 Class C O_6 X	yes	yes	yes	0	0
16	Regression discontinuity	(subsequent to an event, O's are measured; question then asked: did event make a difference?)	no	no	no	0	0

Legend: R (randomization of subjects), O (measurement of dependent variable), X (experimental treatment)

Examples of studies using these designs:

- Design 7: Bessemer (1991). Historical records indicating performance ratings were compared across time from pre-SIMNET to SIMNET condition (author describes as quasi-experiment of transfer using interrupted time series design.).[44]
- Design 8: Whitten, Horey, and Jones (1989). Students were tested at intervals during training on a simulator.
- Design 8 with additional judgment data: Orlansky, Taylor, Levine, and Honig (1997). Cost and training effectiveness evaluation of the MDT2 (Multi-service Distributed Training Testbed), a prototype virtual simulation for training the close air support mission and involving multi-service air and ground forces. Process and outcome measures were obtained on a daily basis during 5-day exercise. Participant judgment data were obtained at end of exercise.

Other examples of studies using these designs are listed in Reference List A-3 in Appendix A.

Test

For purposes of this manual, a test is defined as an experimental trial without requirements for control group, random assignment, or pre- or posttesting and in which performance is measured against a predefined standard. If performance meets the standard, the test is said to be a success. Structurally, a test is equivalent to a 1-shot case study (see Table 3-2, Design 1). However, a 1-shot case study does not necessarily define its success based on meeting a standard. In addition, the test, as used in military training evaluation, is not conducted under the pretense that it is a valid scientific experiment. Tests preclude determining the strength of possible association between X and O, though they may provide weak evidence. If the performance standard is not met, then it may be reasonable to suspect that the experimental treatment is not working—unless the evaluator can infer another cause for the poor performance. If the standard is met, the evaluator cannot conclude that the experimental treatment is the cause—there may be other reasons—although the finding is encouraging.

One justification for conducting tests during training system development is to build confidence or serve as "sanity checks." The test is used to try out an experimental treatment with minimal resources to see if any effect occurs before committing the full resources necessary for a true experiment.[45] In the system developer's world, the most common reason to conduct tests is to determine whether developmental systems are meeting milestone performance requirements; however, test success is usually defined in terms of engineering rather than training performance.

[44] The use of historical versus new data also qualifies this as an ex post facto experiment.

[45] Hiller (1997) wrestled with the dilemma of dealing with questionable data during the development of costly LSTS as follows: "Some data *may* be better than none (only a qualified endorsement for data here, since bad data will mislead). And data collected to disconfirm [a prediction] have far greater utility than randomly collected data. In common parlance, people will typically suggest a "sanity" check when a new [prediction] has been proposed. Any procurement program costing a billion dollars, and possibly critical to the national defense, surely merits a sanity check...." (p. 2.)

Sixteen studies in TCEF were classified as tests based on their methodological descriptions. These appear to fall into two classes: competency tests and estimates of training potential. Competency tests (N=12) evaluate the proficiency of equipment operators and mechanics to perform their jobs to standard after training. Estimates of training potential (N=4) evaluate the potential of developmental training systems to deliver training on predetermined lists of training tasks. These two classes of studies demonstrate how a test can provide useful information while lacking full scientific rigor.

Examples of studies using these designs:

- Competency test: Ennis and Gardner (1990). Soldiers completed knowledge and performance tests and scores were evaluated against a standard. Supporting soldier judgment and SME observational data were gathered. Provides a snapshot of skill and knowledge; there is no direct evidence linking training to performance.
- Estimation of training potential: Smith and Cross (1992). Aircrews performed a variety of individual and collective tasks on simulator and SMEs rated their performance; aircrews also completed questionnaire items.

Other examples of studies using these designs are listed in Reference List A-4 in Appendix A.

Transfer

Transfer experiments attempt to measure the effects of learning in one situation (e.g., using a flight simulator) to performance in another (e.g., flying an aircraft). Transfer can be positive, negative, or indeterminate. Positive transfer is good and negative transfer bad, while indeterminate transfer indicates that training value is unknown.[46] Training evaluators have written many positive things about transfer experiments. For example, Pfeiffer and Browning (1984) state, "There is little doubt that data resulting from carefully designed and well-controlled transfer of training experiments can provide the most convincing evidence of the value of simulators for aircrew training" (p.13). They further comment on the high cost and difficulty of conducting such experiments, concluding that other forms of evaluation must often be substituted.

Twenty-two evaluations in TCEF were classified as transfer experiments based on methodological descriptions. Nine deal with aviation training, 10 with gunnery, and three with training media. These studies cover a narrow a range of subjects—mainly aviation and gunnery. Cost may be a factor because transfer experiments usually cost more than other types of experiments. Military

[46] The amount of transfer is typically computed with transfer effectiveness ratios. See Roscoe (1971, 1972) and Povenmire and Roscoe (1972).

decision-makers have apparently been willing to spend more when evaluating aviation and gunnery training devices and simulators. This willingness may be based on the serious consequences of poor training; that is, aviation accidents or missed targets.

Pfeiffer and Browning describe three classes of transfer experiments, based on purpose:

- Validation: Demonstrate transfer from training device to job.
- Comparison: Compare the amount of transfer from two or more devices to job.
- Relationship: Determine functional relationship between amount of training on device and performance on job.

Table 3-4 illustrates several different types of transfer designs based on Pfeiffer and Browning's framework. Consider first validation studies. Note the similarities between Design 1B in Table 3-4 and Design 1 in Table 3-2, and Design 1A in Table 3-4 and Design 3 in Table 3-2. Is there a difference between these pairs of designs, other than their stated purposes? This is debatable, but it appears that performance measurement in the pre-experiments takes place in a test at a single point in time following training, but in the transfer experiments it occurs over a period of time during a second learning experience, while using actual equipment.[47]

Design 1C is the reverse of Design 1B. Design 1C is a "backward transfer" design, intended to determine the amount of transfer from actual equipment to simulator. Positive backward transfer may imply positive forward transfer, while absence of backward transfer may indicate problems with the design of the simulator.[48] Campbell and Stanley would probably classify Designs 1B and 1C as "pre-experimental," with all that implies.

Consider comparison studies. Designs 2A and 2B in Table 3-4 are identical except that 2A has a control group and 2B does not. Design 2C compares transfer between two different training devices.

Consider relationship studies. Designs 3A and 3B in Table 3-4 are identical except that 3A has a control group and 3B does not.

Of the 23 transfer evaluations, the frequency and percent of studies for different designs are shown in Table 3-4. The relative use of validation, comparison, and relationship studies was comparable. The sample is too small to comment on the breakdown by design type.

[47] If this is true, then Design 1A has a legitimate control group, but it does not appear to rescue Design 1B from the various fatal flaws of 1-shot case studies pointed out by Campbell and Stanley (refer to earlier discussion of pre-experimental designs).

[48] Kaempf (1986) provides the rationale for backward transfer within aviation simulator experiments as follows: "[The] backward transfer paradigm is a relatively low-cost procedure designed to measure the degree to which flying skills transfer from an aircraft to a flight simulator. The paradigm requires that an experienced aviator fly the specified maneuver in the simulator without the benefit of simulator practice. Subjects must meet two criteria. First, the subjects must demonstrate proficiency in the aircraft on the tasks of interest, and second, they must have no experience flying the flight simulator. Backward transfer occurs if the aviator is able to perform the maneuver in the simulator to a desired criterion level of proficiency. Such a finding indicates that positive transfer in the reverse direction, from the simulator to the aircraft, is likely; however, the procedure provides no method to estimate the magnitude of positive transfer. The absence of backward transfer, on the other hand, indicates that the aviators are unable to perform adequately in the flight simulator. Such a finding points to potential problems with either the design or functioning of the flight simulator" (pp. 42-43).

Table 3-4. Three Classes of Transfer Designs (adapted from Pfeiffer & Browning, 1984)

PURPOSE	EXPERIMENT TYPE	DESIGN	GROUPS	DESCRIPTION	FREQ	PER-CENT
Validation	Transfer	1A	E C	SIM-------ACT -------ACT	2	9
	Transfer Quasi-	1B	E	SIM-------ACT	0	0
	Backward Transfer Quasi-	1C	E	ACT-------SIM	4	17
Comparison	Transfer	2A	E1 E2 C	SIM1-------ACT SIM2-------ACT -------ACT	1	4
	Transfer Quasi-	2B	E1 E2	SIM1-------ACT SIM2-------ACT	6	26
	Interdevice Transfer Quasi-	2C	E1 E2	SIM2-------SIM1 SIM1-------SIM2	2	9
Relationship	Transfer	3A	E1 E2 E3 C	SIM-------------ACT SIM----------ACT SIM-------ACT -------ACT	3	13
	Transfer Quasi-	3B	E1 E2 E3	SIM-------------ACT SIM----------ACT SIM-------ACT	4	17

Legend: E, E1, E2, E3 (experimental group); C (control group); SIM (simulator); ACT (actual equipment)

Examples of studies using these designs:

- Design 1A: Browning, McDaniel, Scott, and Smode (1982). A 2-group transfer of training experiment. Both groups were trained on cockpit procedures trainer and SH-3; experimental group also received training on flight simulator (2F64C). Groups were then compared to determine number of flight hours required to reach proficiency.
- Design 1B: None.
- Design 1C: Kaempf (1986). Sixteen instructor pilots who lacked recent experience on a flight simulator performed a set of eight different emergency touchdown maneuvers on a new flight simulator while being graded by SMEs.
- Design 2A: Povenmire and Roscoe (1971). A 4-group transfer experiment: (1) prior flight experience, (2) aircraft only, (3) AN-T-18 simulator, (4) GAT-1 simulator. Students were trained on aircraft only or simulator plus aircraft and their flight performance was later evaluated.
- Design 2B: McDaniel (1987). A 2-group between-subjects transfer experiment. Both groups received classroom training. Group 1 was trained with paper acoustic spectrograms and Group 2 with Passive Acoustic Display Simulator. Both groups were then trained and tested on the Aviation ASW Basic Operator Trainer.

- Design 2C: Smith and Hagman (1993). A 2-group transfer experiment: Group 1 was pretested and trained on MCOFT (Mobile Conduct of Fire Trainer) and posttested on GUARD FIST I (Guard Unit Armory Device, Full Crew Interactive Simulation Trainer); Group 2 did the opposite.
- Design 3A: Hart, Hagman, and Bowne (1990). A 3-group between-subjects transfer experiment: groups (16 subjects in each) received 0, 1, or 3 TOPGUN training sessions and then were tested on COFT (Conduct of Fire Trainer).
- Design 3B: Shute and Gawlick-Grendell (1992). A 2-group between-subjects transfer experiment. Both groups received same training, but differed on amount of practice problems (3 vs. 12 per practice set). Subjects were then tested on a transfer task.

Other examples of studies using these designs are listed in Reference List A-5 in Appendix A.

Ex Post Facto

For purposes of this manual, *ex post facto* "experiments" are defined as studies that use historical data to mimic experiments. The quality of data for these studies depends upon the source. In the best of cases, well-maintained data archives contain data collected in anticipation of ex post facto study. In other situations, the data are whatever an organization has kept. From the former to the latter situation, any ex post facto studies show less planning and are less desirable. Attempting to mimic an experiment with "found" data is risky. Before proceeding, the evaluator should confer with experts in quantitative methods.[49]

Fifteen studies in TCEF were classified as ex post facto based on their methodological descriptions. These appear to fall into two classes: comparison and correlation/regression. Comparison studies (N=11), like 2- or more-group experiments, compare the effects of one or more experimental treatments, but based on historical rather than newly-generated data. Correlation/regression studies (N=4) use one of those statistical methods on historical data to calculate the degree to which a particular type of training contributes to later performance.

[49] Hiller (1994), Boldovici and Bessemer (1994), and Leibrecht (1996) have recommended the use of ex post facto analysis on archival data obtained from the CCTT because of the need to accumulate and integrate data over the long term to separate effects of training from confounding variables. In the case of LSTS such as the CCTT, evaluation with true experiments is impractical but ex post facto analysis may hold the key. This subject is discussed in greater detail later in this manual.

Examples of studies using these designs:

- Comparison: Derrick and Davis (1993). Comparative study of large training system comprising 43 courses taught to pilots, navigators, flight engineers, loadmasters, and maintenance technicians. Study compared the costs and effectiveness of traditional aircrew training system (conducted entirely by USAF) and contractor-delivered (flying training only delivered by USAF). Training folders were examined for the two training programs and training periods to assess training effectiveness. Cost data were obtained by counting resources for both systems; for example, number of graduates, instructors, airplanes, flying hours, training days, overhead staff, types and number of training devices, and so forth
- Correlation: Sterling (1996). Historical gunnery data relating to the use of the BFV platoon gunnery trainer (PGT) and performance during live fire exercises at Grafenwoehr were obtained and correlated. Live-fire performance was positively correlated with increased use of PGT.

Other examples of studies using these designs are listed in Reference List A-6 in Appendix A.

Judgment-Based Evaluations

For purposes of discussion, define judgment-based evaluation as one that relies primarily on human judgment in the form of estimates, ratings, comments, or other expressions to provide training effectiveness data. Judgments are obtained in a structured way.[50] Judgment is a type of data, though it is convenient to refer to as a method. Judgment data may be gathered concurrently with other methods of evaluation, such as experiment, analysis, or survey. When it is, the judgment data are almost always of secondary importance to data gathered with the primary method.

Judgment-based evaluations can be performed on hypothetical ways to train—before a training system exists—provided there is enough descriptive documentation to support analysis. This is an important feature of both judgment-based and analytical evaluations as compared to experiment or survey, both of which usually require existing, functional training systems.

Judgment-Based Evaluation As Experiment

One way to think of judgment-based evaluations is as experiments whose dependent variables are judgment data. This works to a degree. For example, judgment-based evaluations are often used to compare two or more different ways to train; this is analogous to a

[50] Human judgment obtained even in an unstructured way can be powerful. An example is the influence a respected military leader can wield through approval or disapproval.

multi-group experiment.[51] They are used to assess the quality of training programs or the performance of their students; this is analogous to the one-shot case study. Judgment-based evaluations are often used in a more analytical way; for example, to estimate the training potential of a hypothetical training system. This obviously has no experimental equivalent and here the analogy breaks down.

Whose Judgment Is Asked?

Human observers presumed to be knowledgeable about a question being asked provide judgment data. Based on TCEF, these users appear to be of three main types:

- Analysts: Members of the evaluation community who are technically knowledgeable but not SMEs; examples are civilian analysts and test managers.
- Subject-matter experts: Typically, senior and knowledgeable members of the user community, such as master gunners and instructor pilots.
- Users: Typically, the class of individuals whose training is being evaluated, such as students, equipment operators, and crew members.

These categories overlap, but usually apply when describing whose judgment is being asked during an evaluation.

Gathering Judgment Data

Judgment data can be gathered via written or computer-based questionnaire or by interview according to a data collection protocol. The protocol defines the form and content of responses.

Attitudes Versus Technical Estimates

Within TCEF, two general types of judgment data were the most common: attitudes and technical estimates. Attitudinal data express an individual's personal reactions to a training event in terms of likes and dislikes, preferences for or against, suggested improvements, and so forth. Technical estimates express a judgment about a training system or program; an example is estimated effectiveness in providing training on a pre-determined set of training tasks. Technical estimates may also reflect judgment on how well a particular task (such as operating a simulator) was performed.

Attitude data are usually gathered with multiple-choice and open-ended questions and rating scales. Technical estimates are usually gathered with rating scales.

[51] "Experiments" which rely exclusively on judgment data are notorious. Such a study might, for example, compare the classroom use of some novel medium or method based on student reactions. Studies of this nature are fairly common in the educational literature.

Comparative Versus Non-Comparative Evaluations

Judgment data in TCEF were used in one of two ways: comparatively or non-comparatively. In comparative use, an observer states a comparative judgment about two or more alternatives, such as the relative training value of training methods A versus B. In non-comparative use, an observer states a judgment about a single training event, such as the rated quality of a training program. Comparative and non-comparative uses are two of Jeantheau's (1971) four levels of evaluation (qualitative, non-comparative, comparative, and transfer). Although TCEF includes no judgment-based qualitative or transfer evaluations, such evaluations are possible and do occur. A judgment-based qualitative evaluation provides weaker evidence than a non-comparative or comparative evaluation and has limited value. Does judgment-based transfer evaluation provide stronger evidence? It may, if evaluators are able to make valid estimates of transfer. Whether or not this is possible is debatable.

Summary Breakdown

Eighty-seven studies in TCEF used judgment-based evaluation methods. The Overall block in Table 3-5 gives the frequency and percent of use of analyst-, SME-, and user-judgment based evaluations for all evaluations. The Exclusive block shows a breakdown for evaluations in which judgment-based evaluation was the primary evaluation method used (32 evaluations). The Ratio column on the right shows the ratio of Exclusive to Overall. The relative use of judgment by group, from greatest to least, was Analysts (.71), SMEs (.48), and Users (.27). In most cases user judgment was the secondary method used. Conversely, when Analyst or SME judgment data were used, they were usually the primary method. In about one-third of the evaluations where judgment was the primary method used, data were gathered on a hypothetical way to train.

Table 3-5. Frequencies and Percentages of Usage Of Analyst-, SME-, and User Judgment-Based Evaluations in TCEF Sample

JUDGMENT TYPE	OVERALL		EXCLUSIVE		RATIO: EXCLUSIVE/ OVERALL
	FREQ	PERCENT	FREQ	PERCENT	
Analysts	7	8	5	16	0.71
SMEs	25	29	12	38	0.48
Users	55	63	15	47	0.27

Examples of these evaluation methods are described below, and further examples are listed in Appendix A. Without going into these in great depth, it is possible to make a few general observations. First, user-judgment based evaluations are about twice as likely to be used to gather attitude data as technical estimates; this is true whether used alone or in combination with other evaluation methods. SME-based evaluations deal almost exclusively with technical estimates rather than attitudes. By a ratio of about four to one, these estimates express the SME's judgment about the performance of a training system/program or task performance. Analyst-based evaluations appear to be the most rigorous of the judgment-based analyses. Most deal with technical estimates (vs. attitudes) and are non-comparative.

Estimating Training Potential

Many of the technical estimates dealt with the training potential of a developmental or hypothetical system. Participant judgments helped system developers explore new training concepts, test them in prototype form, and support training system development.

Examples of studies using these designs:

- Judgment (Users) (attitudes, non-comparative): Mirabella, Sticha, and Morrison (1997). User reactions to participation in MDT2 training were obtained with a combination of survey questionnaires, group interviews, and observations of training.
- Judgment (Users) (technical estimates, comparative): Thomas, Houck, and Bell (1990). Pilots participated in exercises using a multiplayer aviation simulator and then pilots and controllers rated the value of that training in relation to traditional methods.
- Judgment (SME) (rate human performance, comparative): Quester and Marcus (1984). Supervisors rated performance of students trained in classroom or on the job in 12 occupational categories.
- Judgment (SME) (rate system performance, comparative): Kelly (1995). SMEs separately rated training capabilities of traditional method (Range 400) and Leathernet (pre-build system).
- Judgment (SME) (rate system performance, non-comparative, estimate training potential): Keller, Parrish, Harrison, and Macklin (1992). SMEs separately estimated what tasks could be trained on three alternative aviation simulators.
- Judgment (Analyst) (rate task performance, non-comparative). Kraemer and Bessemer (1987). Analysts closely observed tank crews during SIMNET training, interviewed participants, and inferred effects on live gunnery performance.

Other examples of studies using these designs are listed in Reference Lists A-7 (Users), A-8 (SMEs), and A-9 (Analysts) in Appendix A.

Analytical Evaluations

There is no simple and widely accepted definition of analytical evaluation, although this terminology is commonly used. These evaluations tend to use existing data to evaluate existing or hypothetical ways to train.[52] They are not experiments or surveys and do not use judgment data.[53] So far, this says next to nothing about what they are, but hints at what they have going for them: low cost and the ability to be performed in an office.

For purposes of definition, define analytical evaluation as a method that determines effectiveness based on analytical techniques and using analytical strategies. The dictionary definition of "analysis" refers to separating a whole into component parts and examining the elements and their relations (Merriam-Webster, 1986). Analytical evaluations in TCEF seem to share common steps of problem definition, decomposition into component parts, determining relations among elements, application of logical rules, and generation of conclusions; these are the so-called "analytical techniques." Analytical evaluations are conducted for many different purposes, some more obvious than others. Analytical strategies are ways to dissect, organize, structure, and combine the data for analysis; common strategies are modeling, extrapolation, and task list analysis.

Analytical evaluations can be performed on hypothetical ways to train—before a training system exists—provided there is enough descriptive information to support analysis. In contrast, experiment or survey usually requires functional training systems.

If the definition of analytical evaluation remains vague, it may become clearer in the discussion and concrete examples that follow.

Purposes of Analytical Evaluation

Analytical evaluations in TCEF were conducted for what appear to be three main purposes:[54]

- Evaluate: Assess training effectiveness of a single way to train.
- Compare: Compare the relative effectiveness of two or more ways to train.
- Optimize: Refine the attributes of a training design to maximize its effectiveness.

[52] Recall that "way to train" refers to alternatives such as the use of various instructional media, classroom treatments, and simulations. In analytical evaluations, this is usually some type of simulator or training device.

[53] Analysts must exercise judgment and, insofar as this is done subjectively, it blurs the line between judgment-based and analytic evaluations. In the best of all possible worlds, the analyst follows well-defined procedures that minimize the need to rely on judgment.

[54] Some of the evaluations in TCEF were conducted for additional reasons as well; for example, to investigate the need for a particular new way to train such as a training simulator, or to estimate the training potential of a particular technology. In all cases, these purposes were secondary to one of the three primary purposes listed.

Analytical evaluations may be conducted on existing or hypothetical ways to train.

The three main purposes of evaluation cross with these two ways to train to produce six possibilities (Table 3-6).

Table 3-6. Frequencies and Percentages of Usage of Three Classes of Analysis (Evaluate, Compare, or Optimize) for Existing Versus Hypothetical Systems

CLASS OF ANALYSIS	EXISTING		HYPOTHETICAL		OVERALL	
	FREQ	PERCENT	FREQ	PERCENT	FREQ	PERCENT
Evaluate	13	30	13	30	26	60
Compare	1	2	10	23	11	26
Optimize	0	0	6	14	6	14

Summary Breakdown

Forty-three studies in TCEF used analytical evaluations. In all of these studies, analysis was the primary evaluation method used. Table 3-6 breaks down the frequency and percent of use of analytical evaluations by class of analysis (Evaluate, Compare, or Optimize) and whether for existing or hypothetical training.

Based on these data, it appears that analytical evaluations are about twice as likely to be performed on hypothetical as on existing ways to train. This is roughly the inverse of the ratio for judgment-based analyses. This sample is too small to make comparisons, but suggests that hypothetical ways to train are more often evaluated analytically than based on judgment. The frequency of use of analytical evaluation (Overall) declines from Evaluate (60%) To Compare (26%) To Optimize (14%).

Comparative Versus Non-Comparative Evaluations

The majority of analytical evaluations were non-comparative; approximately one-fourth were comparative. Comparisons were possible where comparative data on two or more alternatives were available. The sample included no analytical evaluations of qualitative or transfer data. The former, though possible in principle, does not make much sense. The latter would be useful, and may be possible, though no such evaluations are present in TCEF.[55]

[55] Analysts have made a number of attempts over the years to develop analytical techniques to predict transfer of training. Two of these methods are *Simulated Transfer* and *FORTE*, discussed later in this chapter.

Examples of studies using these designs:

- Analysis (Evaluate, existing): Simutis, Ward, Harman, Farr, and Kern (1988). Retrospective evaluation and review of a large-scale research program, based on historical data. Review of Army Research Institute (ARI) research 1980-1988 in the basic skills education program (BSEP). Information was gathered from enlisted historical data files and field visits were made to Army posts to observe BSEP training and to and interview participants (administrators, teachers, graduates, supervisors, and commanders). Notes: uses historical data.
- Analysis (Evaluate, hypothetical): McDade (1986). Prospective evaluation of a hypothetical simulator to train BFV (Bradley Fighting Vehicle) drivers. Stated study objective: determine need for driver trainer and if it would be a cost-effective way to train BFV drivers. Driving tasks were identified. Driver training effectiveness was assessed by observing training in schools and units and by questionnaires and interviews with command, supervisors, instructors, and BFV crews. Driver training costs were estimated with and without simulators. Notes: models and evaluates training option.
- Analysis (Compare, existing): Ellis and Parchman (1994). Compared two existing training programs, traditional and CBI-based. The Course Evaluation System (CES) method was used to assess match between course objectives, test items, and instructional presentation for both new (CBI-based) and traditional versions of course. Students completed a questionnaire to assess attention, relevance, confidence, and satisfaction. Test scores were compared between new and old versions of course. Notes: applies checklist evaluation framework.
- Analysis (Compare, hypothetical): Stoloff (1991). Evaluated the relative cost-effectiveness of five different ways of expanding an existing video tele training (VTT) network. Courses suitable for VTT delivery were identified with a selection model and high throughput courses were selected from among these. Costs were then computed for each of the five alternatives and cost-effectiveness was judged by comparing cost of VTT versus that of sending instructors to remote sites. Notes: models and evaluates alternatives.
- Analysis (Optimize, existing): None.
- Analysis (Optimize, hypothetical): Communications Technology Applications, Inc. (1988). Objective of study was to identify an effective training strategy to train soldiers to operate, maintain, and repair JTIDS (Joint Tactical Information Display System), a secure communication system. No precursor training system existed and no training data were available. Training effectiveness forecasting decision analytic framework was developed and applied. Method took into

account estimated task training efficiency, training program effectiveness, and cost data for alternative training strategies. Steps followed: review JTIDS literature, develop training (hypothetical—analyze missions/functions/tasks, generate course structure), analyze training effectiveness, assess trainability, analyze device requirements, determine costs, conduct tradeoff analysis. Notes: uses computer tool.

Other examples of studies using these designs are listed in Reference Lists A-10 (Evaluate), A-11 (Compare), and A-12 (Optimize) in Appendix A.

Analytical Strategies

Analytical strategies are ways to dissect, organize, structure, and combine data for analysis. They are both subtle and obvious. It is probable that many analysts apply strategies without naming them or giving them much thought. However, it is useful to label them for later reference. All strategies are not applicable in all analyses. The choices depend primarily upon the type of data available.

Modeling

Modeling may be the most common strategy. To evaluate, optimize, or compare ways to train, they must be represented in some form. This representation is a model. It may be a given (e.g., a sheet of specifications) or may have to be created (e.g., description of alternative ways to design a distance education system with estimated costs, training effectiveness, etc.). The model may be of something real or hypothetical. The model is to analysis as the actual training system is to an experiment. All of the examples described above use modeling in one form or another. See McDade (1986) and Stoloff (1991).

Analogy

With analogy, apply knowledge about how A (existing) works to predict how B (hypothetical) will work. The hypothetical system is usually a next-generation or close relative of the existing system. Thus, in TCEF, analogy was used to predict CCTT effectiveness based on SIMNET (Noble and Johnson, 1991a,b; Lynn and Palmer, 1991), Breacher effectiveness based on CEV (Skog, Neal, and Fields, 1994); and Heavy Assault Bridge effectiveness based on Breacher (Carroll, 1995). Examples:

- Analogy: Noble and Johnson (1991a,b). Analytical study to determine possible OPTEMPO (Operating Tempo) reductions with adoption of CCTT. CCTT training effectiveness was estimated based on previous analyses of

SIMNET (surrogate system). CCTT training development requirement was examined to determine task areas to be trained; these were compared with task areas covered by SIMNET. Three different training device alternatives were compared (improved SIMNET-T, degraded CCTT, embedded training). Costs were estimated.

- Analogy: Carroll (1995). Objective was to determine the most cost-effective training strategy for Heavy Assault Bridge, a longer version of Breacher. This study extrapolated from the earlier Breacher CTEA. Breacher CTEA was analyzed to identify bridging-specific tasks, and training alternatives were generated; SMEs reviewed these products. Training methods and resources were estimated. Alternative training strategies were developed. Costs were estimated for the alternative strategies. Sensitivity analysis was conducted. Training strategy was determined by comparing relative costs and estimated effectiveness of alternatives.

Extrapolation

Define extrapolation as prediction based on an understanding about how a process works. There are different forms of extrapolation. TCEF includes two: theory and computer model. Though they work differently, both provide a means to predict. For purposes of discussion here, a theory is an explanation of how something works that resolves the various known facts about it and that permits certain predictions. A computer model is a computer program that models a process; given proper inputs, it will predict an outcome. A computer model may be thought of as the instantiation of a theory. Examples:

- Extrapolation from theory: Crawford and Suchan (1996). Analysis conducted to estimate the suitability of various electronic media (e.g., forms of instructional TV, digital video) as substitutes for delivering graduate education to Navy medical officers. Four instructional outcomes based on Gagné and Briggs were identified (know and supply information, apply information within structured situations, exercise judgment in face of uncertainty, understand and change habits of mind). Alternative media were characterized as relatively "lean" or "rich" based upon their ability to support interaction. Expected learning outcomes of the media were estimated based upon their characteristics. The target training modules were then examined based on various criteria to estimate suitability for media. Note: extrapolates effectiveness based on theory.
- Extrapolation from computer model: Muller, Adkins, Belfer, Carter, and Levy (1988). Analysis conducted to determine the most cost-effective of three hypothetical training programs for the NLOS (Non Line of Sight) weapon system: (1) training device intensive, (2) tactical equipment, (3) device/tactical

equipment blend. Hypothetical POIs (Program of Instruction) were designed and then modeled on a "POI optimizer" computer program that takes into account length of time for instruction, media, equipment, student/equipment ratio, type of instructor, student/instructor ratio, set up time, andequipment cost. Notes: models and evaluates alternatives, uses computer tool.

Task List Analysis

Military training is defined in terms of tasks that must be performed, the associated conditions, and standards of performance. Task lists define what is taught in military schools and what military personnel must perform on the job. Hence, tasks are the essential building blocks of training. Task list analysis was used in four studies in TCEF. In all cases it was used to predict how well a training system would be able to support training. These analyses began with lists of tasks on which training was to be conducted. In one case (Burnside, 1990), SMEs then estimated how well the tasks could be performed on a simulator and in the other three cases (Drucker and Campshure, 1990; Fusha, 1989; Thomas and Gainer, 1990, May), personnel attempted to perform the tasks on the simulator and their performance was evaluated. The result in all cases indicated how well the tasks could be performed on the simulator and, in effect, evaluated its potential training effectiveness. Examples:

- Burnside (1990). SMEs rated degree to which selected ARTEP tasks could be performed in SIMNET. Ratings were consolidated with decision rules, reviewed, and coordinated.
- Drucker and Campshure (1990). Analysis conducted to estimate how well SIMNET can be used to train tactical activities conducted during tank platoon operations. The activities performed by armor personnel during combat were identified from field manuals and other documents. The research staff then attempted to perform these activities on SIMNET and recorded estimated fidelity with a checklist.

Historical Data

The evaluator may compile and integrate historical data in the form of an evaluative review. There must be enough data to reach a conclusion about the training value of a specific way to train. Such a review is arguably the operational equivalent of a TEA because it supports conclusions about training effectiveness. However, it will tend to be more narrowly focused, involve less data, and cover a shorter time frame than the typical academic review or meta-analysis.[56] TCEF includes 10 studies that use historical data to reach conclusions. Examples:

[56] This manual largely excludes historical reviews and meta-analyses. A few historical reviews were included on the basis that they focus on very specific training evaluation questions; for example, the utility of basic skills education, CBI, flight simulators, and simulation in marksmanship training.

- Harman (1984). Information was gathered from enlisted historical data files and field visits were made to Army posts to observe BSEP training and survey and interview participants (administrators, teachers, graduates, supervisors, commanders).
- Hall and Rizzo (1975). Survey and review of team training: Research team made site visits to locations where team training was conducted and observed training and interviewed participants. Also conducted literature review. Describes current (1975) team training practices and characteristics (e.g., nature of team performance, coordination, types of training). Recommendations for improving training were developed by comparing practices with literature findings.

Other Analytical Strategies

A considerable amount of research and development work has been conducted to develop analytical methods to analyze and predict training effectiveness. This work has produced an enormous literature that was reviewed by Muckler and Finley (1994a,b). This literature is tangled and confusing, and many of the methods are complex to apply and unvalidated. However, evaluators who need to take an analytical evaluation approach should consider them. Pfeiffer and Horey (1988) summarize and briefly consider the strengths and weaknesses of 18 different analytical evaluation methods.[57] Some methods that may be useful are Simulated Transfer, FORTE (FORecasting Training Effectiveness), and Comparison-Based Prediction (CBP). These methods apply in evaluating training devices and simulators. In evaluating schoolhouse training, consider the Instructional Quality Inventory. Readers interested in further information on these and the other methods should consult Pfeiffer and Horey (1988). Also, refer to these sources: FORTE (Pfeiffer, Evans, and Ford, 1985) and CBP (Klein, Johns, Perez, and Mirabella, 1985). See Chapter 6 for more information on these methods.

One of the purposes of analytical evaluations is optimization—to refine the attributes of a training design to maximize its effectiveness. The TRADOC Analysis Center, White Sands Missile Range (TRAC-WSMR), has developed an analytical method that can be used for this purpose by evaluating the cost and training effectiveness of various mixes of field training and training using training aids, devices, simulators, and simulations. The "training mix model" is a computer program that incorporates the expected cost of acquiring and using training systems with their expected effectiveness in terms of ability to train required tasks (Djang, Butler, Laferriere, and Hughes, 1993). TRAC-WSMR is continuing to develop, apply, and refine this method. See Chapter 6 for more information.

Available procedural guidance on analytical methods is discussed further in Chapter 6.

[57] One of these, *Instructional Quality Inventory*, was applied in Ellis and Parchman (1994), described above. Another, *Task Commonality Analysis*, is very similar to the *task list analysis* strategy described in the text.

Surveys

For purposes of discussion, define survey as a process to gather data from a group presumed to be knowledgeable about a training issue. The scale of the survey may range from small to large. Data may be gathered in a variety of different ways, including questionnaire, interview, and observation (Bouchard, 1976; Fowler, 1993). Surveys commonly use judgment data. Hence, Judgment and Survey methods overlap. What distinguishes them is scale: surveys are larger than judgment-based evaluations.

Surveys have long been used to answer training effectiveness questions. Within TCEF, large-scale surveys were used when investigators had to gather such a large amount of data that it was the only practical method. Small-scale surveys are used to (1) assess the status of training programs and (2) investigate the application of new technologies in the field.

Summary Breakdown

Fourteen studies in TCEF used survey methodology. This number is misleading because seven of these studies are in fact separate volumes of the Army Training Study (Brown [1978a,b,c,d,e,f,g]), a large survey conducted for the Army. The *Army Training Study* and the *Combat Effective Training Management Study* (Rosenblum, 1979) are both large-scale surveys conducted in the post-Vietnam era to evaluate Army training at a systems level. Four of the studies were small-scale surveys conducted to assess the status of training programs. Two small-scale surveys were conducted to investigate the application of new technologies in the field.

Comparative Versus Non-Comparative Evaluations

All of the TCEF surveys were qualitative or non-comparative. It is conceivable that survey respondents could provide comparative or transfer data, if asked the right questions; examples of questions include "compare system A with system B" or "estimate the degree of transfer from device A to device B."

Examples of studies using these designs:

- Large-scale survey: Brown (1978). Seven-volume survey. Study group applied ARTS model (describes what Army training is, ought to be, and should do and defines the objectives of training) to evaluate Army training; also conducted field surveys and collected data at various Army posts and schools. Comprehensive review of Army training,
- Small-scale survey (assess training program status): Hall and Rizzo (1975). Survey team made site visits to locations where team training was conducted and observed training and interviewed participants. Also conducted literature review. Describes current (1975) team training practices and characteristics (e.g., nature of team performance, coordination, types of training). Recommendations for improving training were developed by comparing practices with literature findings.
- Small-scale survey (investigate application of new technology in field): Pugh, Parchman, and Simpson, H. (1991). Field survey was conducted among representative sample of ITV sites in public education, industry, and military. Data were gathered via observation and interview.

Other examples of studies using these designs are listed in Reference List A-13 in Appendix A.

4 CASE STUDIES

Procedural guidance (see Chapters 3, 6, and 7) presents the principles, procedures, and theory of training evaluation. This is essential information, but often lacks the context that makes real-world evaluation problems unique. Another way to learn how to evaluate training is to study cases; that is, concrete examples of how evaluations have been conducted in the past. Cases provide insight into evaluators' decision-making, problem-solving strategies, evaluation methods, reporting, lessons learned, and general practices. They may show what was done well and poorly, what mistakes were made, and where the risks lie in the future. Cases provide vicarious experience that theory cannot. Good cases illustrate good evaluation practice. However, even flawed cases are useful if they help evaluators avoid future errors.

This chapter makes case studies of SIMNET/CCTT and MDT2. It describes the evaluation of these two cases in terms of individual evaluation events, methods used, and the evaluation process followed.

Two reference lists at the end of this chapter contain complete citations for publications cited in this chapter. See Reference Lists 4-1 (SIMNET/CCTT) and 4-2 (MDT2).

Finding Cases

It would be nice if training system developers documented every aspect of their training evaluations and archived them in a readily-accessible source such as the Defense Technical Information Center (DTIC). The developer of a new training system could then search DTIC, identify relevant prior evaluations, and use them as models in evaluating a new system. The evaluator of a new constructive simulation such as the JSIMS (Joint SIMulation System) would likely look at earlier-generation simulations of the same type, such as the Aggregate Level Simulation Protocol (ALSP) Confederation of Models, BBS (Brigade/Battalion Battle Simulation), or CBS (Corps Battle Simulation). The evaluator of new virtual or advanced distributed simulations would likely look to SIMNET, CCTT, CATT (Combined Arms Tactical Trainer), or the Navy's BFTT (Battle Force Tactical Trainer). Unfortunately, the number of training evaluations conducted and published on most of these systems is few.

While developing this manual, an attempt was made to obtain prior training evaluation studies for all of these systems. The results were disappointing. For many cases of interest, few published evaluations were available from DTIC. It is not clear whether this reflects lack of (a) training evaluation, (b) publication, or (c) both. Evaluators of new systems may have to go to system developers or proponents to track down all relevant evaluations.

Two Good Cases

The search was successful for two cases: SIMNET/CCTT,[58] and MDT2. These are both virtual simulations. Both SIMNET and MDT2 received good R&D funding and became the focus of research interest because of their innovativeness. Developed with DARPA (Defense Advanced Research Projects Agency) support, SIMNET was adopted into the Army in 1989, and in 1990 the Army began procurement of its production follow-on, the CCTT (Alluisi, 1991). SIMNET consists of simulators of tactical command posts, M1 Abrams tanks, M2/M3 Bradley fighting vehicles, Army helicopters, and fixed-wing close air support aircraft linked into a network that allows crews in different locations to train together on a common battlefield. Because it arose under DARPA, SIMNET did not undergo the usual DoD development process for LSTS in terms of reporting, milestone testing, and so forth. SIMNET is associated with dozens of different training evaluations, yet none is definitive. Reviewers attempting to determine SIMNET's training effectiveness of have had to rely on the weight of evidence from many small evaluations rather than a single conclusive one. CCTT is undergoing the usual development and operational testing.

The MDT2 is also unique. First, it was an experimental testbed for advanced simulation concepts and never intended to become an operational system. It was born, existed briefly, and retired. While operational it, too, received much research interest. Second, researchers attempted to determine its overall training effectiveness in terms of a range of dependent variables. It may be the best single case study of how to evaluate LSTS.

SIMNET/CCTT

The TCEF includes nearly 50 publications on SIMNET/CCT[59] (Reference List 4-1). These are believed to be most of the significant publications that are readily available on the subject.[60] Approximately two-thirds of these publications are evaluations. The rest fit into the categories of reviews, evaluation methodology, evaluation plans, or evaluation tools.

[58] CCTT is the follow-on to SIMNET. Though different systems, their developmental history shares much in common and they are treated together.

[59] SIMNET nicely illustrates the point made in Chapter 2 about how evaluations of LSTS may unfold over a period of years and and consist of several different evaluations, at different points in time, and often for different purposes.

[60] This sample does not include unpublished or classified evaluation reports or conference or journal publications that did not receive wide circulation.

Reviews

The various reviews of SIMNET are of interest to evaluators because they provide information on the system development process. This process is difficult to understand based on individual evaluation reports, which provide only snapshot coverage; the reviews provide a moving picture from start to finish.

Alluisi (1991) provides a historical review of SIMNET/CCTT development, from the first work at DARPA in 1983 on through its various tests, evolution into CCTT, and status circa 1991. It shows where SIMNET came from and how it got to be the way it is. SIMNET was developed iteratively, using rapid prototyping and quick modification, in a risk-tolerant environment.

Cosby (1995) provides an engaging first-person historical account of the origin of the SIMNET concept and its evolution into an actual training system. Cosby identifies the people who made SIMNET a reality and speculates on how simulation technology will influence training in the future.

For reasons of historical interest, readers may want to examine the initial design study for SIMNET (Gurwitz, Burke, Calvin, Chatterjee, and Harris, 1983), which describes the SIMNET concept, hardware and software components, and hypothetical exercises that might be run with the system as envisioned before it existed.

Three more recent reviews have attempted to estimate the training and cost-effectiveness of simulation technology. In each case the reviews have considered SIMNET and examined the evidence available to date to make a judgment. As there is no single definitive evaluation of SIMNET/CCTT, these reviews present and weigh the available evidence. See Angier, Alluisi, and Horowitz (1992); Orlansky, Dahlman, Hammon, Metzko, Taylor, and Youngblut (1994); and Worley, Simpson, Moses, Aylward, Bailey, and Fish (1996).

Evaluation Methodology

The challenges of evaluating LSTS have caused some evaluators to reconsider evaluation methodology and to recommend new or modified approaches. Two excellent papers and an unpublished memorandum on this subject have had a strong influence on the evaluation framework presented in Chapter 8 and in the discussion of evaluation problems areas in Chapter 5. The contents of these publications are sketched below. Chapters 5 and 8 consider the issues they raise.

Boldovici and Bessemer (1994) critically review several SIMNET evaluations, describe their shortcomings, and suggest alternative evaluation strategies for future evaluations.[61] They highlight the common shortcomings of field studies and offer suggestions for alternative, innovative research methods; for example, in-device learning experiments, quasi-transfer experiments, correlational analyses, and quasi-experiments.

Boldovici and Kolasinski (1997) describe three techniques (hypothesis tests, power analysis, confidence intervals) to apply in designing experiments comparing two or more alternative treatments to enable the findings to declare the treatments equivalent. This cannot be done with the types of statistics commonly used in such instances, as the finding of "no significant difference" does not prove equivalence.

Hiller (1994, 7 February) wrote an issue paper describing a proposed approach to evaluating CCTT for its milestone III decision. He makes the case that traditional experimental design cannot estimate effects of CCTT on readiness and proposes a two-aspect evaluation strategy: (1) long-term data collection from units training with and without SIMNET/CCTT and (2) experimental applications of CCTT.

Evaluation Plans

Evaluation plans may be useful as models in creating new plans and for the particular methods, data collection instruments, and procedures they describe. TCEF includes three such plans.

Clapper and Schwab (1986) present a plan to test capabilities of SIMNET to support platoon-level command and control exercises and to train individual tasks. Eight tank platoons are to be evaluated on a tactical pretest with actual equipment. Four platoons train on SIMNET (experimental group) and four on standard exercises (control group). SMEs evaluate performance of both groups on posttest (same as pretest).

Smith (1989) presents a plan to prospectively evaluate CCTT based on its ability to train companies/teams on particular ARTEP collective tasks as judged by SMEs, using SIMNET as surrogate for CCTT.

TEXCOM (1998) presents a plan to evaluate CCTT using a 2-group experiment, where the groups consist of task-force sized elements. Group 1 receives training on CCTT and then goes to NTC. Group 2 goes directly to NTC without using CCTT. Performance of both forces is evaluated at NTC and compared.

[61] Among their reviews are several evaluations described below: Kraemer and Bessemer (1987); Schwab and Gound (1988); Brown, Pishel, and Southard (1988); TEXCOM (1990); Burnside (1990); and Drucker and Campshure (1990).

Evaluation Tools

SIMNET/CCTT evaluators have published descriptions of evaluation tools that may be useful to others. The tools include data collection methods and a task, conditions, and standards data base.

Meliza and Tan (1992) provide a framework for evaluating unit performance data during SIMNET exercises following the model used at NTC. They provide guidance to use SIMNET UPAS (Unit Performance Assessment System) to collect and analyze unit performance data. UPAS collects, filters, and analyzes data broadcast over the network, loads data into a relational data base, integrates data with terrain and planning information, and provides graphic and tabular displays. Meliza, Bessemer, and Tan (1992) describe UPAS development. Meliza, Bessemer, Burnside, and Shlechter (1992) describe platoon-level AAR aids for use with UPAS. Meliza (1993) is a SIMNET collective training data base manual that describes an Army repository of collective task, conditions, and standards information.

Evaluations

Table 4-1 summarizes SIMNET/CCTT evaluations in TCEF by author and evaluation method and submethod used. Evaluations are listed in order of year of publication within each Method row. Most of these evaluations either (a) evaluate some aspect of SIMNET/CCTT or (b) estimate SIMNET/CCTT training potential. Some of the evaluations are more remotely linked. Holstead (1989), Crane, and Berger (1993), and Thomas and Gainer (1990, May) investigated the utility of a SIMNET-type simulation for aviation training. Hartley, Quillinan, and Kruse (1990a,b) and Watson (1992) deal with computer models. Hoffman (1997) describes the introduction of a virtual training program within an actual unit. And Bessemer and Myers (1998) evaluate a structured, simulation-based training program.

The evaluations are summarized briefly below. Reference List 4-1 (SIMNET/CCTT) at the end of this chapter and References at the end of this manual contain complete citations for publications cited in this chapter.

Experiments

Schwab and Gound (1988). A 2-group experiment to evaluate SIMNET's capability to support platoon-level command and control exercises to train individual and collective tasks. Groups were pretested, received SIMNET or field training, and were posttested in a field exercise. Dependent variables were STX (Situational Training Exercise) GO scores.

Table 4-1. SIMNET/CCTT Evaluations by Authors and Evaluation Method and Submethod

METHOD	AUTHOR (YEAR)	SUBMETHOD
Experiment	Schwab & Gound (1988)	experiment (true): pretest-posttest control group
	Brown, Pishel, & Southard (1988)	experiment (true): pretest-posttest control group
	TEXCOM (1990)	experiment (pre-): 1-group pretest-posttest
	Smith & Graham (1990)	experiment (true): postest-only control group
	Hartley, Quillinan, & Kruse (1990a,b)	experiment (test)
	Shlechter, Bessemer, & Kolosh (1991)	experiment (ex post facto)
	Bessemer (1991)	experiment (ex post facto), experiment (quasi-)
	Watson (1992)	experiment (true): postest-only control group
	Smith & Cross (1992)	experiment (test)
	Shlechter, Bessemer, Nesselroade, & Anthony (1995)	experiment (quasi-): equivalent time samples design
	TEXCOM (1997)	experiment (pre-): 1-group pretest-posttest design
Analysis	Fusha (1989)	analysis (evaluate)
	Drucker & Campshure (1990)	analysis (evaluate)
	Burnside (1990)	analysis (evaluate)
	Thomas & Gainer (1990, May)	analysis (evaluate)
	Noble & Johnson (1991a,b)	analysis (compare)
	Lynn & Palmer (1991).	analysis (evaluate)
	Scott, Djang, & Laferriere (1995)	analysis (optimize)
	Finley (1997)	analysis (evaluate)
Judgment	Kraemer & Bessemer (1987)	judgment (analysts)
	Brown & Mullis (1988a,b)	judgment (users)
	Holstead (1989)	judgment (SMEs)
	Crane & Berger (1993)	judgment (users)
	Hoffman (1997)	judgment (users)
	Bessemer & Myers (1998)	judgment (analysts)
Survey	Fletcher (1988)	survey

Brown, Pishel, and Southard (1988). A 2-group experiment to compare SIMNET and field training. Eight platoons were pretested, four each participated in SIMNET and field training, then all participated in ARTEPs. Dependent variables were platoon performance, command and control, and leadership.

TEXCOM (1990). Test to estimate training potential of CCTT using SIMNET as surrogate. Platoons were pretested on actual vehicles, underwent SIMNET training, and were posttested on actual vehicles. Dependent variables were various tactical indicators, such as exchange ratio, percent losses by force, and shots/kill.

Smith and Graham (1990). Evaluated the use of SIMNET as a soldier evaluation device by comparing soldier performance on field and SIMNET tests using the multirait-multimethod matrix and analysis of variance technique. Total of 120 tank crews participated in field tests and comparable tests in SIMNET and performance in both were compared on four dimensions: (1) command and control, (2) communications, (3) position location, (4) combat driving.

Hartley, Quillinan, and Kruse (1990a,b). Describes the process followed in verifying and validating direct fire and direct/indirect vulnerability models in SIMNET-T for M-1 main gun and M-2 25mm. In each case, behavior of SIMNET-T was compared with baseline mathematical models for the simulated effects.

Shlechter, Bessemer, and Kolosh (1991). Evaluated the benefits gained by students acting in the role of platoon leaders during SIMNET training in armor officer basic school. Regression analyses were used to determine how students demonstrated leadership skills as compared to their peers who played non-leadership roles.

Bessemer (1991). Ex post facto, quasi-experimental assessment of transfer of SIMNET training to student officer performance in field training. Performance ratings were compared across time from pre-SIMNET to SIMNET condition.

Watson (1992). Study compared the tactical outcomes of SIMNET-D and Janus-T using the same scenario.

Smith and Cross (1992). Study to (a) assess experienced crew members' ability to perform selected individual and collective tasks in AIRNET and (b) identify the specific design attributes that make it difficult for crew members to perform tasks to standards in AIRNET. Aircrews performed a variety of individual and collective tasks on simulator and SMEs rated their performance.

Shlechter, Bessemer, Nesselroade, and Anthony (1995). Study to evaluate training effectiveness of Reserve Component Virtual Training Program's (RCVTP) simulator-based training program. Unit performance scores on training tables were obtained and compared across six successive training tables.

TEXCOM (1997). Test to evaluate the training transfer capability of SIMNET. Platoons were pretested on actual vehicles, underwent SIMNET training, and were posttested on actual vehicles; SIMNET was used as surrogate for CCTT.

Analytical Evaluations

Fusha (1989). Analytical assessment of the potential utility of SIMNET to support training on Bradley operations and tasks at platoon and squad level. Study group evaluated mission training plans, estimated whether or not tasks and drills could be trained on SIMNET, and developed task lists and scenarios to evaluate the trainable tasks and data collection instruments to assess trainability. SMEs attempted to execute the scenarios on SIMNET and completed questionnaires regarding trainability.

Drucker and Campshure (1990). An analysis to estimate how well SIMNET can be used to train tactical activities conducted during tank platoon operations. The activities performed by armor personnel during combat were identified from field manuals and other documents. The research staff then attempted to perform these activities on SIMNET and recorded estimated fidelity with a checklist. Study estimated training potential using task list analysis.

Burnside (1990). SMEs rated degree to which selected ARTEP tasks could be performed in SIMNET. Ratings were consolidated with decision rules, reviewed, and coordinated. Study estimated training potential using task list analysis.

Thomas and Gainer (1990, May). Case study to evaluate how well AIRNET could be used to train ARTEP tasks. Tasks were selected. Pilots used AIRNET to conduct simulated missions. SMEs rated their performance and AIRNET performance for each task. Subjects completed questionnaires about technical performance of system.

Noble and Johnson (1991a,b). Analytical study to determine possible OPTEMPO reductions with adoption of CCTT. CCTT training effectiveness was estimated based on previous analyses of SIMNET: CCTT TDR (Training Device Requirement) was examined to determine task areas to be trained; these were compared with task areas covered by SIMNET. Three different training device alternatives were compared (improved SIMNET-T, degraded CCTT, embedded training). Costs were estimated.

Lynn and Palmer (1991). Analytical evaluation of hypothetical training. Analysts reviewed various CCTT conceptual documents (Concept Evaluation Program, Training Device Needs Statement, Training Device Requirement, System Specification) and reports (reliability, force development test and experimentation final report) and estimated operational effectiveness of CCTT. CCTT strengths and weaknesses were extrapolated from those of SIMNET.

Scott, Djang, and Laferriere (1995). Objective was to find best way to field future CCTT into reserves. Reserve soldiers with CCTT experience rated effectiveness of current training; ratings provided estimates of best training mission scenarios. Mathematical models were use to estimate costs of three fielding alternatives.

Finley (1997). Prospective evaluation of the capability of CCTT to provide a suitable environment for training involving degraded communications. Analyses were performed to identify training needs in armor and mechanized infantry units using single channel ground/air radio systems. Capabilities of initial CCTT to simulate realistic variations in communications quality were then estimated.

Judgment-Based Evaluations

Kraemer and Bessemer (1987). Judgment-based evaluation of hypothetical training. Analysts closely observed tank crews during SIMNET training, interviewed participants, and inferred effects on live gunnery performance.

Brown and Mullis (1988a). Assessment of soldier perceptions of the relative fidelity of physical, visual, and aural characteristics of SIMNET. Total of 26 tank crewmen were trained on SIMNET, used it for a while, and then rated its realism and value for training

Brown and Mullis (1988b). Assessment of soldier perceptions about using SIMNET training in preparation for the Canadian Army Trophy (CAT) competition. Total of 145 tank crewmen were trained on SIMNET, used it for a while, and then rated its realism and value for training

Holstead (1989). Large-scale operational effectiveness appraisal of SIMNET to assess the capability/potential of SIMNET to train Air Force personnel. SMEs participated in SIMNET CAS (Close Air Support) exercises and rated capability of SIMNET to provide training on tactical aviation tasks.

Crane and Berger (1993). Judgment-based assessment of utility of SIMNET-compatible air combat simulator for training pilots using simulated combat exercises. Pilots participated in exercises using a multiplayer aviation simulator and then rated desirability of receiving additional training.

Hoffman (1997). Describes the introduction of simulator-based virtual training program within an actual unit, and the process used to identify and resolve problems. This is not a formal evaluation, but illustrates how trainers may introduce an innovation and work through problems. Hoffman apparently identified problems on site. Participants completed questionnaires.

Bessemer and Myers (1998). Evaluation of training program using structured, simulation-based training and description of process for monitoring similar programs. Army Research Institute teams observed initial implementation of the program, noted problems, and discussed with site contractor. Evaluators later reviewed literature and consulted with TQM experts to derive evaluation methodology, consisting of steps for organizing, identifying problems, developing indicators, monitoring processes, and developing and adopting changes.

Survey

Fletcher (1988). Field survey conducted during early SIMNET implementation to get reactions of participants. Commanders and crews at all levels gave ratings and comments regarding SIMNET performance, how well it exercised different skills, its appropriate training role, and acceptance.

Evaluation Overview

These evaluations occurred over a period of years and demonstrate that evaluation is a process, not an event.

The evaluations were conducted for several different reasons. The most important was to satisfy military milestone requirements. Some evaluations estimated SIMNET/CCTT training potential—and some compared the effects of SIMNET/CCTT training with traditional training. These last two reasons are different sides of the same coin, separated by time. The first asks the question, "How well can it train?" The second asks, "How well does it train?" These evaluations were conducted for most of the reasons discussed in Chapter 1, although it was not obvious if any was used to improve SIMNET/CCTT design.

The evaluations relied first on experiment, second on analysis, and third on judgment as the evaluation method. One survey was conducted.

The experiments used several different dependent variables; for example, tactical performance, leadership performance, command and control and communications, and gunnery performance. Many also gathered supporting judgment data from participants. These variables primarily reflect training outcomes in terms of combat skills.

MDT2

The TCEF includes nearly 20 publications on the MDT2 (Reference List 4-2). These include a definitive technical report (Orlansky, Taylor, Levine, and Honig, 1997) and conference paper (Taylor, Orlansky, Levine, Honig, and Moses, 1996), a survey of user reactions (Mirabella, Sticha, and Morrison, 1997), and lessons learned (Bell, Dwyer, Love, Meliza, Mirabella, and Moses, 1997a,b). Other reports and papers describe data collection methods and instruments, network hardware and software, and the MDT2 evaluation from perspectives of different participants. The literature on the MDT2 project is ample for use in designing a new LSTS evaluation.

The following description of the MDT2 project is based closely on Orlansky et al. and Bell et al.

Purpose

The MDT2 project was conducted during 1994 and 1995 to test the feasibility of using virtual simulation to conduct multi-service training on the CAS mission. The work was supported primarily by funding from the DMSO (Defense Modeling and Simulation Office) and conducted by a consortium of researchers from Army, Navy, Air Force, and IDA (Institute for Defense Analyses). In addition, OSD (Office of the Secretary of Defense) funded IDA to analyze the cost-effectiveness of an operational version of MDT2. The project brought together researchers from different Services to evaluate the use of virtual simulation for multi-service training.

Simulators

MDT2 connected 11 simulators at four locations around the United States, enabling them to interact in real time to conduct simulated combat exercises against simulated enemy forces. Simulators, services, types of participants, and locations are shown in Table 4-2. Participants included Army, Marine Corps, and Air Force units. Eight different types of simulators representing enemy and friendly forces were linked to conduct exercises modeled on those at the NTC.

Exercises

A total of 19 personnel participated in each battalion task force exercise against a regiment-sized enemy represented by semi-automated forces. Two sets of exercises were conducted, each over a period of five days. The first set of exercises was conducted 23-26 May 1994 and the second set 13-17 February 1995.

The first day was used for familiarization. Defensive exercises were conducted on the second and fourth days and offensive exercises on the third and fifth days. The exercises followed scenarios. The missions required integration of CAS with the fire and movement of an armored battalion task force that was part of an Army brigade attached to a Marine expeditionary force. An airborne Marine forward air controller in an OV-10 observation aircraft, a Marine laser designator team with a ground forward air controller, and an Air Force tactical air control party and F-16 attack pilots supported CAS. Several of the CAS missions used laser-guided bombs dropped by the F-16s on enemy targets designated by the Marine laser designator team.

Table 4-2. Simulators, Service, Types of Participant, and Locations Linked in MDT2 (adapted from Orlansky et al., 1997).

SIMULATORS	SERVICE	TYPES OF PARTICIPANTS	LOCATION
• Tactical Operations Center (TOC) • M1 Abrams tanks • M2 Bradley fighting vehicles	Army	• Key staff members to include Tactical Air Control Party • Command Group to include Air Liaison Officer of Enlisted Terminal Attack Controller • Task Force Scouts • Company Commander and Exec	• Mounted Warfare Test Bed, Ft. Knox, KY
• Semi-Automated Forces (SAFOR) (enemy and friendly)	Army		
• F-16 aircraft simulator	Air Force	• Attack Pilots	• Armstrong Laboratory, Mesa, AZ
• Deployed Forward Observer/Module Unit Laser Equipment (DFO/MULE) Laser Target Designator	Marine Corps	• Ground Forward Air Controller and Laser Team	
• Helmet mounted display simulator • OV-10 aircraft simulator	Air Force Marine Corps	• Airborne Forward Air Controller	• Naval Air Warfare Center, Patuxent River, MD
• Recording and observation			• Institute for Defense Analyses, Alexandria, VA • Armstrong Laboratory, Mesa, AZ • Mounted Warfare Test Bed, Ft. Knox, KY

Method

The MDT2 evaluation used experiment and survey methods. During each exercise, subjects attempted to perform their tasks in support of the CAS mission. Process and outcome measures were obtained during each exercise. Process measures reflect activities occurring during training. Outcome (or product) measures reflect the products or consequences of training.

An AAR was conducted at the conclusion of each exercise (Moses, 1995). After each battle, trainers and O/Cs (Observer/Controller) shared data, observations, and comments to provide feedback to participants. An hour or so later, all personnel were linked in a teleconference to discuss the battle and view replays.

Because a new exercise was completed each day, it was possible to gather data comparing performance on successive days.

A survey was conducted at the end of the week.

The MDT2 evaluation has been characterized in some MDT2 publications as a case study. The experimental design resembles Campbell and Stanley's Design 1 (one-shot case study); that is, on any day of the study, a single group underwent an experimental treatment and was evaluated, with no control group, pretest, or random assignment. However, the fact that this process was repeated five times during the week, enabling comparison of performance across successive days, makes it akin to one of the quasi-experimental designs; for example, Campbell and Stanley Design 8 (equivalent time samples design).

Dependent Variables

Process Measures

Two types of process measures were collected during exercises: TARGETs (Targeted Acceptable Response to Generated Events or Tasks), and TOMs (Teamwork Observation Measure). SMEs generated both measures by observing participants and recording their observations using special data collection protocols.

The TARGETs methodology is described as follows in Fowlkes, Lane, Salas, Franz, and Oser (1994):

> It is a form of structured observation in which (a) task events are introduced to provide opportunities for teams to demonstrate specific team-related behaviors; (b) acceptable team responses to each of the events are determined a priori by utilizing team task analyses, subject-matter experts, and so forth; and (c) the appropriate responses to events are scored as either present or absent (p. 47).

Fowlkes et al. state that the TARGETs methodology is relatively easy to apply, does not require SMEs, and possesses high inter-rater reliability. For further discussion of TARGETs, refer to Dwyer, Fowlkes, Oser, Salas, and Lane (1997) and Fowlkes, Dwyer, Oser, and Salas (1997).

Teamwork Observation Measure data reflect the adequacy of interactions among team members for each of three mission phases (planning, contact point, attack) and four dimensions (communication, coordination, situational awareness, adaptability). The data identify strengths and weaknesses in teamwork (Dwyer, Fowlkes, Oser, and Salas, 1996). The teamwork dimensions were divided into subdimensions for each phase of the exercise (Table 4-3). For example, the communication dimension was broken down by correct format; proper terminology; clear, concise, and accurate; and acknowledgments.

Table 4-3. Dimensions and Subdimensions of TOM (Teamwork Observation Measure) (adapted from Orlansky et al., 1997)

PHASE	COMMUNI-CATION	TEAM COORDINATION	SITUATIONAL AWARENESS	TEAM ADAPTABILITY
Planning or Control Point or Attack	• Correct format • Proper terminology • Clear, concise, and accurate • Acknowledgments • Other	• Synchronized actions • Timely passing of information • Familiar with others' jobs • Other	• Maintained "big picture" • Identified potential problem areas • Aware of resources available • Provided information in advance • Other	• Backup plans • Smooth transition to backup plans • Quickly adjusted to situational changes • Other

Outcome Measures

The Army developed UPAS to calculate and display performance measures and summary statistics for SIMNET exercises. Its development and operation are described in Meliza, Bessemer, Burnside, and Shlechter (1992), Meliza (1993), and Meliza, Bessemer, and Tan (1992). The UPAS gathers data from five sources (network, terrain, unit plans, radio communications, and direct observations) and generates information on vehicle appearance, status, and status change and fire, indirect fire, and impact. The UPAS data were recorded during each exercise, permitting later playback to develop these outcome measures:

- Number, timing, and frequency of bombs released by F-16s
- Number of vehicles hit, damaged, or destroyed
- Percentage of bombs resulting in a vehicle impact or near impact
- Number of bombs causing damage or destruction
- Timing and volume of artillery direct fires and CAS fires
- Timing and location of direct and supporting fire impacts

Survey

On the final Friday at the end of both the 1994 and 1995 demonstrations, all participants and O/Cs completed a written survey to give their judgments, opinions, and comments on how well MDT2 worked and what value it added. This survey is described in Mirabella, Sticha, and Morrison (1997); see also Mirabella (1995).

Findings

For purposes here, study findings are less important than the way the data were organized for analysis and presentation. The evaluators displayed the data on particular dependent variables in the form of learning curves so that changes in performance could be readily detected. This study lacked a control group as may be the case in many future LSTS evaluations. However, being able to discern skill growth (or lack thereof) adds a dimension in judging the training value of a new system.

The following briefly excerpts portions of the findings as reported in Orlansky et al.; Dwyer, Oser, and Fowlkes (1995); and Dwyer, Fowlkes, Oser, and Salas (1996).

Process Measures

Data were broken down by exercise phase (planning, contact point, attack). Figure 4-1 shows the TARGETs data for the planning phase for the functions target selection, airspace coordination areas, control of aircraft, synchronization, and overall performance. Each data point represents the mean percent correct for all O/Cs across all CAS missions for the day. The shape of the curves shows that performance improved during the week.

Figure 4-2 shows corresponding TOMs data for the functions communication, coordination, situational awareness, adaptability, and overall performance. Each data point represents the mean rating given by the O/Cs. Again, it is obvious that performance improved during the week.

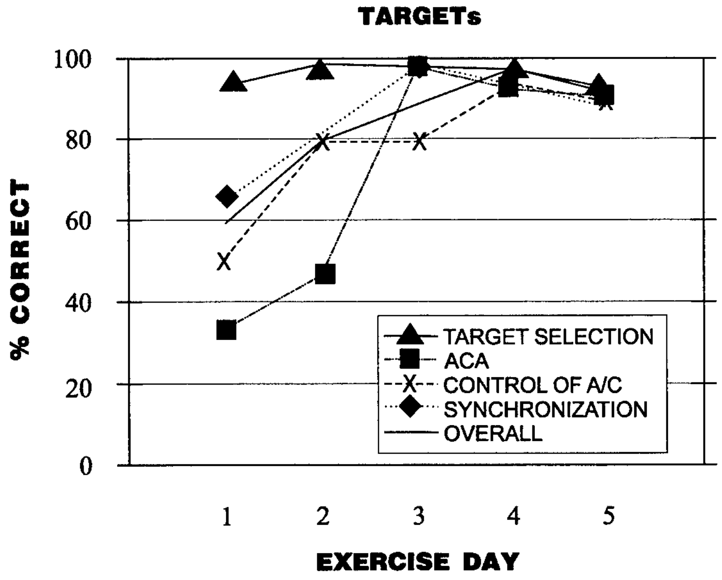


Figure 4-1. TARGETs data for planning phase (from Orlansky et al., 1997)

Refer to the three references cited above for complete data.

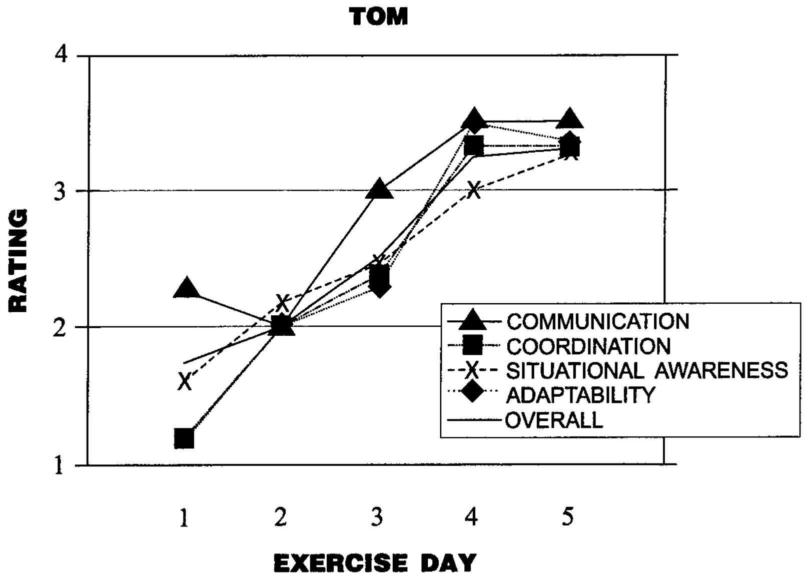


Figure 4-2. TOM data for planning phase (from Orlansky et al., 1997)

Outcome Measures

Table 4-4 shows bombing performance on successive days for the 1995 exercise. Table 4-5 shows CAS kills, misses, and average engagement times for that exercise. Again, refer to the sources for complete data.

Table 4-4. Bombing Performance for February 1995 Exercise (from Orlansky et al., 1997).

	EXERCISE DAY				
PERFORMANCE MEASURE	**1**	**2**	**3**	**4**	**5**
Number of bombs released per day	4	10	11	19	14
Mean releases per mission	1.0	2.5	3.7	3.8	4.7
% of missions with 3 or more releases	0	50	66	80	100
Mean time between releases (minutes: seconds)	18:31	7:12	3:39	4:15	4:12
Number of releases separated by less than 1 minute	0	2	5	11	10

Table 4-5. CAS Kills, Misses, and Average Engagement Time for February 1995 Exercise (adapted from Orlansky et al., 1997)

	EXERCISE DAY			
PERFORMANCE MEASURE	**2 (Defense)**	**3 (Offense)**	**4 (Defense)**	**5 (Offense)**
CAS Kills	3	5	7	9
CAS Misses	3	5	5	2
Average Engagement Time (min.)	4	3	2	1.5

Table 4-6. Results of MDT2 Survey (from Orlansky et al., 1997)

ISSUE	SURVEY ITEM (31 subjects each year, across all sites)	% AGREE	
		1994	1995
Need	The opportunity provided by MDT2 to practice with personnel from other services is necessary for training CAS	90	90
	MDT2 is a good training system for CAS because it focuses on critical training needs	90	74
	Given the chance, I would like to train with the MDT2 on a periodic basis	94	83
Credibility	MDT2 can be an effective trainer for CAS with only a few minor modifications	81	55
	A positive aspect of MDT2 is that it gives more realistic feedback on CAS kills than in field exercises or at combat training centers	94	69
	I can apply more realistic CAS tactics in MDT2 than I can in field exercises or at combat training centers	77	53
Multi-Service Value	Experience on MDT2 will make me better able to interact with members of other services to plan for and execute CAS missions in combat	90	90
	Training with MDT2 will give me a better understanding of the jobs and role or personnel from other services in planning and conducting CAS	84	87
Role in Training Cycle	Experience with MDT2 will better prepare me for field exercises on CAS mission, such as those at Air Warrior and NTC	87	90
	Training on MDT2 can supplement service-specific CAS training	87	77
Expected Impact	The training that MDT2 provides can be applied directly to combat	97	100
	Estimate the extent to which your experience with MDT2 has affected your ability to perform your role in a mission that uses CAS	93	94

Survey

Table 4-6 gives a few of the results of surveys for both 1994 and 1995. While it might be impractical to gather such data on a daily basis, being able to compare between two successive sets of exercises adds an extra dimension to the data. The survey is described in Mirabella, Sticha, and Morrison (1997).

Cost Analysis

Cost analyses are described in Orlansky et al. and Taylor et al. Two cost analyses were conducted. The first estimated the costs of developing and operating the MDT2 for training exercises. The second compared the relative costs of conducting a one-week MDT2 training simulation with a one-week field exercise. In simple terms, the first estimate attempted to answer the question: "How much did MDT2 cost?" The second attempted to answer the question: "What is the relative cost of training CAS in MDT2 or field exercises?"

The cost analyses in this study are exemplary models of this type of analysis.

Lessons Learned

In 1997, Bell et al. published their recommendations for planning and conducting multi-service training with virtual simulation. These appear to have grown out of the MDT2 project experience. The authors discuss the goals of multi-service training, principles of virtual simulation, designing and planning training exercises, exercise preparation and execution, archiving, and post-exercise training review. Before publication, this report was reviewed by most of the key players in the MDT2 project. Presumably, the lessons it offers are a reasonable expression of what that group learned. These lessons are helpful to anyone evaluating a large-scale training simulation.

Evaluation Overview

This evaluation contrasts with that of SIMNET/CCTT in a number of different ways. The most obvious is that MDT2 was evaluated in two one-week events whereas the SIMNET/CCTT evaluation was an ongoing process lasting more than a decade. SIMNET/CCTT is a real system and MDT2 a testbed; MDT2 evaluators were free to focus their attention on the training effectiveness of their creation without being preoccupied with the need to pass milestone tests or transition their system into the real world.

The MDT2 evaluation was well enough conceived and conducted for evaluators to draw meaningful conclusions about MDT2 training effectiveness and cost. The evidence amassed in the many SIMNET/CCTT evaluations is persuasive of that system's effectiveness. However, no single evaluation of that system presents as strong a case as the MDT2 evaluation.

The evaluation methods and range of dependent variables all contributed to the authority of the MDT2 evaluation. TARGETs, TOMs, and UPAS process measures provided insight into how well the MDT2 functioned. Outcome measures demonstrated that it was effective for training. The repeated measures gathered on successive days enabled the generation of learning curves that demonstrated steady improvement throughout the exercises.

The MDT2 evaluation lacked important elements of a traditional laboratory experiment; for example, random assignment, control group, and pre- and posttesting. Collecting performance data on successive days compensates somewhat for the lack of pretest and posttest. In field tests, random assignment is seldom possible. Control groups represent a special problem.

Additional Topics

Some additional topics, not covered in the text, that may interest the reader, are:

- Engineering the MDT2 network: see Bell (1996), Bell (1995); Rakolta (1994), Loral Systems (1994)
- Hardware and software lessons learned: see Colburn, Farrow, and McDonough (1994)

Reference List 4-1. SIMNET/CCTT Publications

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33*(3), 343-362.

Angier, B.N., Alluisi, E.A., Horowitz, S.A. (1992). *Simulators and Enhanced Training.* IDA Paper P-2672. Institute for Defense Analyses, Alexandria, VA, 1992.

Bessemer, D.W. & Myers, W.E. (1998). *Sustaining and improving structured simulation-based training.* RR 1722. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA344895)

Bessemer, D.W. (1991). *Transfer of SIMNET training in the armor officer basic course.* TR 920. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA233198)

Boldovici, J.A. & Bessemer, D.W. (1994). *Training research with distributed interactive simulation: Lessons learned from simulation networking.* TR 1006. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285584)

Boldovici, J. A. & Kolasinski, E. M. (1997). How to make decisions about the effectiveness of device-based training: Elaborations on what everybody knows. *Military Psychology, 9,* 121-135.

Brown, R. & Mullis, C. (1988a). *Simulation networking assessment of perceptions - I.* TRASANA Report No. LR-1-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB118627)

Brown, R. & Mullis, C. (1988b). *Simulation networking assessment of perceptions - II.* TRASANA Report No. LR-2-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB146645)

Brown, R.E., Pishel, R.G., & Southard L.D. (1988). *Simulation networking (SIMNET) preliminary training developments study.* TRAC-WSMR-TEA-8-99. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB120874)

Burnside, B.L. (1990). *Assessing the capabilities of training simulations: A method and simulation network (SIMNET) application.* ARI RR 1565. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226354)

Clapper, D. & Schwab, J. (1986). *Concept evaluation program of simulation networking (SIMNET).* Ft. Knox, KY: U.S. Army Armor and Engineering Board. (ADB114371)

Cosby, N.L. (1995). *SIMNET: An insider's perspective.* IDA Document D1661. Alexandria, VA: Institute for Defense Analyses. (ADA294786)

Crane, P.M. & Berger, S.C. (1993). *Multiplayer simulator based training for air combat.* Williams AFB, AZ: Air Force Armstrong Laboratory.

Drucker, E.H. & Campshure, D.A. (1990). *An analysis of tank platoon operations and their simulation on simulation networking (SIMNET).* RP 90-22. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA017009)

Finley, D.L. (1997). *Simulation-based communications realism and platoon training in the close combat tactical trainer (CCTT).* TR 1064. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337692)

Fletcher, J.D. (1988). *Responses of the 1/10 cavalry to SIMNET.* IDA Analysis Memorandum No. M-494. Arlington, VA: Defense Sciences Office. (ADA200499).

Fusha, J.E. (1989). *Simulation networking (SIMNET): Evaluation of institutional/ USAIS (U. S. Army Infantry School) use of SIMNET-T. Phases 1 and 2.* RN 2-89. Ft. Benning, GA: U.S. Army Infantry School. (ADA137722)

Gurwitz, R., Burke, E., Calvin, J., Chatterjee, A., & Harris, M. (1983). *Large-scale simulation network design study.* Bolt, Beraneck, & Newman. (ADA134662)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990a). *Verification and Validation of SIMNET-T. Phase 1. K/DSRD-116.* Oak Ridge, TN: Martin Marietta Energy Systems, Inc. (ADB147354)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990b). *Verification and validation of SIMNET-T. K/DSRD-117.* Oak Ridge, TN: Martin Marietta Energy Systems, Inc. (ADB147355)

Hiller, J.W. (1994, 7 February). *Close combat tactical trainer (CCTT) evaluation planning.* Memorandum to COL Thomas A. Horton. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hoffman, G.R. (1997). *Combat support and combat service support expansion to the virtual training program SIMNET battalion exercise: History and lessons learned.* RR 1717. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA341201)

Holstead, J.R. (1989). *Large scale simulation networking (SIMNET) operational effectiveness appraisal (OEA) final report.* TAC Project 89-190T. Eglin AFB, FL: U.S.A.F. Tactical Air Warfare Center. (ADB133378)

Kraemer, R.E. & Bessemer, D.W. (1987). *U.S. tank platoon training for the 1987 Canadian army trophy (CAT) competition using a simulation networking (SIMNET) system.* RR 1457. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA191076)

Lynn, J. & Palmer, K.L. (1991). *Independent operational assessment of the close combat tactical trainer (CCTT).* OA-0200. Alexandria, VA: U.S. Army Operational Test and Evaluation Command. (ADB160088)

Meliza, L.L. & Tan, S.C. (1992). *SIMNET unit performance assessment system (UPAS) version 2.5 user's guide.* ARI-RP-96-05. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA318046)

Meliza, L.L. (1993). *Simulation networking/training requirements relational database: User's guide.* RP 94-01. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275634)

Meliza, L.L., Bessemer, D.W., Burnside, B.L., & Shlechter, T.M. (1992). *Platoon-level after action review aids in the SIMNET unit performance assessment system (UPAS).* TR 956. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA254909)

Meliza, L.L., Bessemer, D.W., & Tan, S.C. (1992). *Unit performance assessment systems development.* TR 1008. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285805)

Noble, J.L. & Johnson D.R. (1991a). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 1: Executive summary.* TRAC-WSMR-CTEA-91-018-1. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157064)

Noble, J.L. & Johnson D.R. (1991b). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 2: Main report.* TRAC-WSMR-CTEA-91-018-2. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB173567)

Orlansky, J., Dahlman, C.J., Hammon, C.P., Metzko, J., Taylor, H.L., & Youngblut, C. (1994). *The value of simulation for training.* IDA Paper P-2982. Alexandria, VA: Institute for Defense Analyses. (ADA289174)

Schwab, J. & Gound, D. (1988). *Concept evaluation program of simulation networking (SIMNET).* Ft. Knox, KY: U.S. Army Armor and Engineering Board. (ADB120711)

Scott, B.B., Djang, P., Laferriere, R. (1995). *Reserve component (RC) mobile close combat tactical trainer (M-CCTT) integration and deployment study.* TRAC-WSMR-TR-95-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Center. (ADB202913)

Shlechter, T.M. Bessemer, D.W., & Kolosh, K.P. (1991). *The effects of SIMNET role-playing on the training of prospective platoon leaders.* TR 938. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA244913)

Shlechter, T.M., Bessemer, D.W., Nesselroade, P., & Anthony, J. (1995). *An initial evaluation of a simulation-based training program for Army National Guard units.* ARI RR 1679. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA297271)

Smith, B.W., & Cross, K.D. (1992). *Assessment of Army Aviators' Ability to Perform Individual and Collective Tasks in the Aviation Networked Simulator (AIRNET).* RN-92-32. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA250293)

Smith, J.W. (1989). *The operational independent evaluation plan (IEP) for the close combat tactical trainer (CCTT) force development testing and experimentation (FDTE).* Ft. Leavenworth, KS: TRADOC Independent Evaluation Directorate. (ADB129761)

Smith, S.E. & Graham, S.E. (1990). *Comparability of an armor field and simulation networking (SIMNET) performance test.* TR 895. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226353)

TEXCOM (1990). *Close Combat Tactical Trainer (CCTT). Force Development Testing and Experimentation (FDTE).* TCATC-FD-0200. Ft. Hood, TX: Author. (ADB147145)

TEXCOM (1997). *Test data report for the Close Combat Tactical Trainer limited user test.* TDR-97-LUT-1645A. Ft. Hood, TX: Author. (ADB228904)

TEXCOM (1998). *Initial operational test and evaluation: Close Combat Tactical Trainer (CCTT) event design plan.* Ft. Hood, TX: Author. (ADB233901)

Thomas, B.W. & Gainer, C.A. (1990, May). Simulation networking: Low fidelity simulation in U.S. Army aviation. *Proceedings of the Royal Aeronautical Society* (pp. 18.1-18.11). London, England.

Watson, B.L. (1992). *SIMNET-D/JANUS(T) comparison study.* TRAC-WSMR-TM-92-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Command. (ADB164784)

Worley, R.D., Simpson, H.K., Moses, F.L., Aylward, M., Bailey, M., & Fish, D. (1996). *Utility of modeling and simulation in the department of defense: Initial data collection.* IDA Document D-1825. Alexandria, VA: Institute for Defense Analyses. (ADA312153)

Reference List 4-2. MDT2 Publications

Bell, H. (1995). Symposium on distributed simulation for military training of teams/groups: The engineering of a training network. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting,* San Diego, CA, 1311-1315.

Bell, H. (1996). Panel of Multi-Service Distributed Training Testbed: DIS training of military teams/groups: The engineering of a training network. *Proceedings of the 17th. I/ITSEC Conference.* Albuquerque, NM, 365-370.

Bell, H.H., Dwyer, D.J., Love, J.F., Meliza, L.L., Mirabella, & Moses, F.L. (1997a). *Recommendations for planning and conducting multi-service tactical training with distributed interactive simulation technology.* ARI RP-97-03. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328480)

Bell, H.H., Dwyer, D.J., Love, J.F., Meliza, L.L., Mirabella, & Moses, F.L. (1997b). *Recommendations for planning and conducting multi-service tactical training with distributed interactive simulation technology: Appendices.* RP 97-04. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA336275)

Colburn, E. Farrow, S., & McDonough, J. (1994). *ADST multi-service distributed training testbed (MDT2) lessons learned.* ADST/WDL/TR-94-W003312. Orlando, FL: Loral Systems ADST Program Office. (ADA282380)

Dwyer, D. J., Oser, R.L., & Fowlkes, J.E. (1995). Symposium on distributed simulation for military training of teams/groups: A case study of distributed training and training performance. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, 1316-1320.

Dwyer, D.J., Fowlkes, J., Oser, R.L., & Salas, E. (1996). Panel on multi-service distributed training testbed, DIS training of military teams/groups: Case study results using distributed interactive simulation for close air support. *Proceedings of the 1996 International Training Equipment Conference.* The Hague, Netherlands, 371-380.

Dwyer, D.J., Fowlkes, J.E., Oser, R.L., Salas, E., & Lane, N.E. (1997). Team performance measurement in distributed environments: the TARGETs methodology. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team Performance Assessment and Measurement: Theory, Methods and Applications* (pp. 137-153). Hillsdale, NJ: Lawrence Erlbaum.

Fowlkes, J., Dwyer, D.J., Oser, R.L., & Salas, E. (1997). Event-based approach to training. *Proceedings of the 19th. I/ITSEC Conference.* Albuquerque, NM.

Fowlkes, J.E., Lane, N.E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETS methodology. *Military Psychology. 6,* 47-61.

Loral Systems (1994). *Protocol extensions to DIS and interface requirements specification (IRS) for the multi-service distributed training testbed (MDT2). Revision 1.0.* ADST/WDL/TR--94-W003412. Orlando, FL: Author. (ADB198789)

Meliza, L.L. (1993). *Simulation networking/training requirements relational database: User's guide.* RP 94-01. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275634)

Meliza, L.L., Bessemer, D.W., & Tan, S.C. (1992). *Unit performance assessment systems development.* TR 1008. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285805)

Meliza, L.L., Bessemer, D.W., Burnside, B.L., & Shlechter, T.M. (1992). *Platoon-level after action review aids in the SIMNET unit performance assessment system (UPAS).* TR 956. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA254909)

Mirabella, A. (1995). Symposium on distributed simulation for military training of teams/groups: MDT2 system assessment and effectiveness. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, 1321-1325.

Mirabella, A., Sticha, P., & Morrison, J. (1997). *Assessment of user reactions to the multi-service distributed training testbed (MDT2) system.* ARI TR 1061. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328473)

Moses, F. L. (1995). Symposium on distributed simulation for military training of teams/groups: The challenge of distributed training. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting,* San Diego, CA, 1306-1310.

Orlansky, J., Taylor, H.L., Levine, D.B., & Honig, J.G. (1997). *The cost and effectiveness of the multi-service distributed training testbed (MDT2) for training close air support.* IDA Paper P-3284. Alexandria, VA: Institute for Defense Analyses.

Rakolta, M. J. (1994). *Network analysis of the multi-service distributed training testbed (MDT2) wide area network. Revision 1.0.* ADST/WDL/TR--94-W003419. Orlando, FL: Loral Systems ADST Program Office. (ADB198826)

Taylor, H.L., Orlansky, J. Levine, D.B., Honig, J.G., & Moses, F.L. (1996). Evaluation of the performance and cost-effectiveness of the multi-service distributed training testbed (MDT2). Royal Aeronautical Society. *Conference Proceedings: The Progress and Direction of Distributed Interactive Simulation,* November 6-7, 1996.

5 EVALUATION PROBLEM AREAS

This chapter identifies and discusses problems commonly encountered when evaluating military training. Military training evaluations are usually field evaluations that impose restrictions that laboratory evaluations do not. Methodological choices can also cause problems. Problems arise if the evaluator uses an inappropriate evaluation method, chooses inadequate dependent variables, or does not collect high quality data. In some cases the solution to a problem is obvious, but in others it is not. The procedural guidance in Chapter 6 may help where the solution is not obvious. The sources cited in this chapter and in Chapter 6 provide additional information on the subject.

The chapter is organized in three sections: field versus laboratory evaluations, lessons learned, and critiques of field evaluation practice.

Field Versus Laboratory Evaluations

The "field" is the normal operating environment of military personnel and their equipment. This might be on board ship, within a military unit, in a classroom, or somewhere else that troops operate. The "laboratory" is an artificial setting in which evaluators exercise a high degree of control over extraneous variables. The distinction between laboratory and field is not as much one of geography as of the degree of control the evaluator can exercise. Evaluators exercise relatively more control over "laboratory" studies and relatively less over "field" studies, regardless of the actual physical setting. Most military training evaluations are field evaluations.

Time Frames, Unfolding of Events, and Subjects

Evaluators usually distinguish between laboratory and field evaluations. For example, Bouchard (1976) identifies several special characteristics and difficulties of field settings: in the field, independent variables may show greater intensity and range, studies may occur over longer time intervals and according to a natural rather than artificial unfolding of events, and the various treatments may be more varied. Among the special difficulties of field evaluations are causal ambiguity, limitations on subjects, and the cost and time of conducting the field study.

In practical terms, with all of the things going on in a field evaluation...with all of the different interested parties and their varied agendas...with the high cost of using military personnel and equipment...with the inevitable reporting deadlines hovering over everything...with all of the careers hanging on the line and in expectation of certain outcomes, well—field evaluations can be difficult and messy.

Field Evaluation Constraints

Johnson and Baker (1974) describe several important differences between laboratory and field tests[62] in terms that suggest that field tests are inferior, limited by constraints that compromise them relative to the laboratory: field tests address real but messy problems, are time- and resource-limited, usually lack full control, have multiple objectives, and have a criterion problem. They state, "Field research is often considered to be a 'dirty' version of the laboratory research paradigm even by its proponents" (p. 208.). One problem that evaluators frequently face is that they become involved in a system development late:

> Frequently, [evaluators] are not involved in this early stage of system development. Only later are they drawn into this hotbed of disillusionment and frustration, usually to disprove (or prove) the other guy's point of view.... The moral for field research is simple: Get involved as early in the system's life as possible. (p. 205.)

System Complexity

Large-scale training simulations usually include training aids, devices, stand-alone simulators, and, for higher echelon units (e.g., squadron, wing, battalion, brigade, division, corps, theater army, etc.), simulations driven by computer-based mathematical models. These days, most forms of training short of actual combat involve simulation. If the simulation is simple (e.g., a device to train target search and detection or a part task simulator for tank driving), then evaluation can be simple. Unfortunately, evaluating the effectiveness of LSTS is difficult because of their complexity. Training programs for LSTS seldom rely exclusively on a single simulation. Instead, they use a mix of training devices, stand-alone simulators, field training with real weapon systems and equipment, and the LSTS (Hiller, 1998, 6 August). Depending upon training echelon, more than one type of LSTS may be used; for example, constructive and virtual. Apart from simulation quality, the training program in which the simulation is used limits training effectiveness.[63] The measurement of training effectiveness is also influenced by the management and performance measurement and feedback mechanisms available (Hiller, 1994, 7 February).

[62] The authors refer to "tests" rather than experiments. Within the lexicon used in the present manual, a test is considered to be a type of experiment and hence these ideas generalize to experiments as well as tests.

[63] For example, ARI's SIMNET training program for armor battalions covers only a fraction of the battalion's missions (Defend in Sector, Movement to Contact, and Deliberate Attack) and contains over 100 detailed lesson plans, each at three levels of difficulty (crawl, walk, run) (Campbell, Campbell, Sanders, Flynn, and Myers, 1995). Judging training effectiveness based on this small sample of missions is risky.

Developing System As a Moving Target

The developing system is a moving target, constantly changing as it evolves, and posing an evaluation challenge that changes across time. During the planning stages, the evaluator must work not with an actual system but with its description, specifications, and other paper representations. Only later does the actual system begin to take shape. Johnson and Baker note: "As various major subsystems and components emerge during...system design, they become candidates for test and evaluation. The results of these subsystem tests are fed back into the ongoing development process"[64] (p. 206). The more the system evolves the greater the precision of evaluation data.

Relevance and Cost-Effectiveness

Despite the problems faced in field evaluations, Johnson and Baker stress the importance of such evaluations because of their real world relevance and contribution to assuring that systems developed are cost-effective.

External and Internal Validity

Field experiments are generally acknowledged to have greater external validity than laboratory experiments. In an early paper considering alternative research designs for experiments in field settings, Campbell (1957) compared designs based on internal validity (can you predict the outcome based on the treatment?) and external validity (does the outcome generalize to other populations, settings, and variables?). He argued that the controls required to assure internal validity often jeopardize representativeness; that is, external validity. A controlled laboratory experiment may predict outcomes in the laboratory, but the constraints of that experiment may prevent the effect from generalizing to the world at large.

The internal/external validity tradeoff is important to military training evaluators, who want to apply their research findings to the real world. Military decision-makers are reluctant to risk their forces to test in battle technological innovations whose only proof of effectiveness has been demonstrated under laboratory conditions. This bias toward field tests is evident in DoD and Service acquisition regulations, which stress the importance of field testing (Simpson, 1995).

[64] This "take" on evaluation—as a process that supports development to improve a design—is consistent with that presented in Chapter 2. The idea is old.

A Philosophy of Laboratory and Field Research

In a methodological paper on the development of intelligent tutoring systems (ITS), Shute and Regian (1993) articulate a philosophy that makes use of both laboratory and field research at different points in time so as to capitalize on the strengths of each:

> Our approach to managing the tradeoff between internal and external validity is to begin with laboratory research (high experimental control and internal validity) and slowly increase external validity, ultimately studying the intervention in the target instructional context (field research). We believe that neither laboratory nor field research alone will give a complete and accurate picture of the instructional effectiveness of a particular intervention (p. 247).

The idea is illustrated by Figure 5-1, a notional relationship between internal and external validity for laboratory and field studies. This simplification helps structure thinking about the tradeoff between laboratory and field research and the strengths and weaknesses of each.

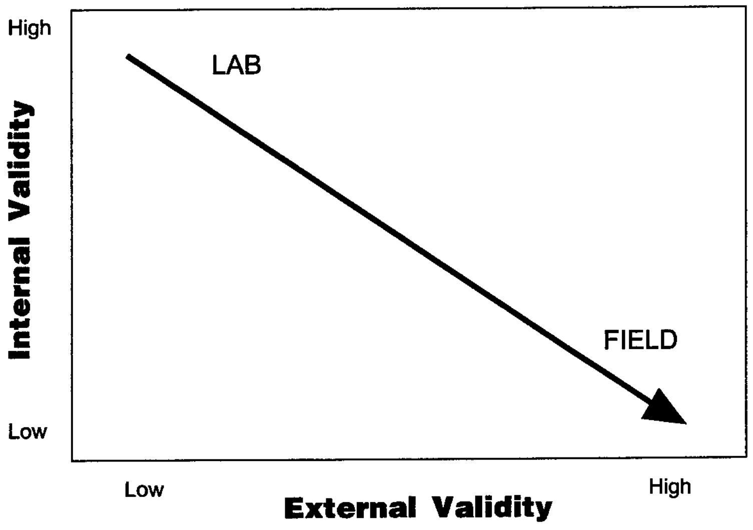


Figure 5-1. Notational representation of a simple inverse relationship between internal and external validity for laboratory and field studies (adapted from Shute and Regian, 1993).

A Contentious Debate

Some methodological purists advocate laboratory research and condemn methods used by field researchers. In their defense, the field researchers argue that they do the best they can under the circumstances and that a higher authority—such as a military command—restricts their ability to select the number of subjects needed, make random assignments, control for certain extraneous variables, and so forth. How does the evaluator get past this debate?[65]

It is reasonable to acknowledge it and move on. Argument will flare whenever purists of differing views are in the same room or on the same Internet, for that matter. Let the debate be waged. Meanwhile, there is a job to be done. The work is important. Consider the point made by Johnson and Baker about the importance of conducting these evaluations. The task is difficult but someone must perform it. Additionally, evaluators often make a sharper distinction between laboratory and field research than necessary. Good research can be conducted in the field—and bad research in the laboratory.

True experiments—conducted in laboratory or field—aspire to the ideal of the laboratory experiment. True experiments often can be conducted in the field, although this is more difficult than in the laboratory. Table 2-4 shows that more than one-third of the TCEF training evaluations involving experiments were classified as true experiments. Considering how much bad has been written about pre-experiments, it is surprising how infrequently they were used.

The evaluator has a professional obligation. Within limits, the evaluator can influence how an evaluation is conducted. He or she should conduct the best possible evaluation under real-world circumstances without (a) *a priori* giving up the game or (b) losing credibility (and any chance at the game) by asking for what military decision-makers cannot give. The title of Johnson and Baker's article is *Field Testing: The Delicate Compromise*, in acknowledgment of the fact that, in conducting field evaluations, compromise is inevitable. Knowing where to draw this line is a matter of professional judgment.

Lessons Learned

System developers and evaluators have documented lessons learned and provided recommendations for future evaluators. If a development and evaluation is uneventful, there is no point in writing up lessons learned. The lessons are usually based on "learning experiences," in which mistakes were made, analyzed, and written up so that future evaluators may avoid them. Lessons learned reflect the perspectives of their authors. In one of the cases below, different groups compiled different sets of lessons learned

[65] The argument can become heated. The author once observed a proponent of laboratory research accuse a peer who advocated a more flexible position of being an apologist for "junk science" whose doubtful results would put at risk the lives of young combatants, possibly offspring and relatives of the audience.

for the same evaluation.[66] The differences can be interesting. The evaluator will have to decide how well lessons learned in one development apply in another.

Alluisi (1991) offers a set of lessons learned about the SIMNET/CCTT development. Bell, Dwyer, Love, Meliza, Mirabella, and Moses (1997a) offer the training evaluator's perspective on MDT2. Colburn, Farrow, and McDonough offer the system contractor's perspective on MDT2. Solick and Lussier (1988) offer lessons for conducting command and staff training with constructive simulation.

Alluisi's lessons are less about evaluation than about what is needed to assure the success of a new development:

1. address recognized, real, and substantial needs
2. with realistic objectives
3. using feasible enabling technologies
4. applied in iterative, rapid prototyping, innovative approaches
5. that make frequent use of concrete demonstrations
6. with customer participation and high-level customer support in a risk-tolerant research •and development environment
7. with competent people
8. organized into a development team with appropriate leadership (p. 359)

Most of these recommendations are obvious. For example, who in this time of limited budgets would advocate developing a system that does not address real needs, with realistic objectives, using feasible technologies? Who would argue with the need for competent people and leadership? However, in the midst of these are recommendations 4, 5, and 6, which take controversial positions on rapid prototyping, frequent use of concrete demonstrations, risk-tolerance, and customer participation.

Bell et al. provide a cookbook for planning, conducting, and evaluating a virtual simulation exercise based on the MDT2 experience. In outline, they recommend that evaluators take these steps:[67]

- Assign responsibilities
- Schedule participants, sites, network, O/Cs
- Identify
  - Training objectives
  - Functional requirements
  - Scenarios and mission
  - Assessment metrics
  - Feedback/AAR
  - Exercise preparation and execution
  - Exercise management requirements
  - Communication requirements

[66] For example, compare the lessons learned by the MDT2 training evaluators (Bell, Dwyer, Love, Meliza, Mirabella, and Moses, 1997a) with those of the Loral contractor team (Colburn, Farrow, and McDonough, 1994).

[67] These are believed to be the most important steps and to convey the spirit of what Bell et al. recommended. Several steps and substeps were deleted for the sake of brevity. Refer to the source for a complete description.

- Ready site
- Ready exercise
- Conduct exercise
- Archive exercise baseline and exercise data

The outline leaves the impression that conducting an exercise involves many small details that must be coordinated according to plan by management with the support of a team of technical specialists working a variety of technical areas. Management provides oversight. The project plan integrates training requirements with engineering. A multidisciplinary team is needed consisting of training developers, human performance experts, network engineers, site representatives, SMEs, O/Cs, and representatives of the training audience. The plan makes less of evaluation than of constructing and implementing the exercise.

Colburn, Farrow, and McDonough were members of the Loral contractor team that provided the MDT2 simulation. Their stated goal was not to evaluate MDT2 but to provide information to potential future users of the simulation. The lessons they provide deal largely with management control and coordination of sites, activities, and schedules; and hardware and software considerations necessary for a successful simulation. They also emphasize the importance of proper preparation of simulation participants, and maximizing the capabilities and utilization of tools for AAR. Evidently, it was their experience that some troops using the simulation were ill-prepared and that the UPAS data available for use in AAR were underutilized. They state: "The key to proper utilization of the system is to provide a timeframe for trainers, engineers, and members of the performance measurement team to meet and discuss the tools that are available...." These words suggest that there was a disconnect between contractors and trainers. This is reinforced by other comments indicating that scenarios used in exercises were developed without consultation with the engineering team or chief trainer. Subsequently, some of the exercises could not be performed. In concluding remarks, the authors comment, "All the components of the exercise should be viewed as part of one system and effort to integrate and employ them should be led by one individual, as project leader." This recommendation and other comments made suggest that the contractor felt that there was inadequate management oversight.

If there was disenchantment with exercise management, it might be because the lessons emerged from developmental systems; that is, SIMNET/CCTT and MDT2. MDT2 was a particularly rough case. It existed only briefly and the trainers, simulation participants, contractors, and others involved in its demonstration had little time to create, implement, use, or evaluate its training. It is reasonable to infer that those involved in this project were harried by the challenge of fitting all the pieces together and making them work in a relatively short period of time. This may explain the emphasis on planning and management; these offer a path through the maze of problems inherent in developing a new system.

By contrast, Solick and Lussier offer the perspective of 10 years of research on command and staff training with automated battle simulations, and are less preoccupied with the mechanics of pulling off the simulation than with making it work effectively for training. Their findings were that the simulations have (a) excessive staff requirements, (b) lack system support for scenario development, (c) lack system control over information and intelligence and (d) lack performance measurement capabilities. Their recommendations mirror these findings. Among other things, they recommend developing a systemic model to minimize support requirements and using on-line data capturing techniques for performance measurement.

While Solick and Lussier's findings are interesting, they do not share much in common with those of the other studies cited in this section. One way to compare and contrast them is by using the metaphor of the automobile. Solick and Lussier's vehicle is old but reliable. Despite poor fuel mileage and other chronic shortcomings, they are confident that it will start and get them to their destination. This cannot be said of the MDT2. Its users must assemble their vehicle from scratch each time they use it. While they would like to focus their full attention on its use for training, they cannot do this until they take care of all the technical details to assure that it will function properly.

If there is a lesson here, it is that training evaluations require attention to the basics (making the simulation work in the narrowest sense) before they can be used to evaluate training. Early on, it would appear, an inordinate amount of energy must be spent working out all the details of the simulation, defining training requirements, training participants, and so forth. Only after attending to these matters is it possible to evaluate training.

Moreover, if the first thing isn't done first, the second is impossible. The training evaluator cannot safely ignore the overall management of the project or leave the details to others. The evaluator need not be manager, but must be close enough to management to wield an influence.

Critiques of Field Evaluation Practice

Researchers sometimes critique the evaluation methods used by others. Critique may be incidental, as in a passing comment in a report or article. Sometimes the critique is an important part of what a study is about. Papers whose declared subject is methodology often critique status quo as prelude to whatever innovation the author endorses. Such critiques are revealing, although they sometimes read as debates in progress.[68]

[68] The methodological papers discussed in Chapter 2 make reasonable but somewhat conflicting recommendations on how to evaluate LSTS and are open to debate. These papers include Hiller (1997); Bell, Dwyer, Love, Meliza, Mirabella, & Moses (1997a, b); Garlinger and Fallesen (1988); and Alluisi (1991). None of these papers offers an unchallenged critique of existing evaluation practice or formula for future practice.

An important exception is Boldovici and Bessemer's (1994) *Training Research with Distributed Interactive Simulation: Lessons Learned from Simulation Networking*, a critique of SIMNET evaluation practice. They focus on several different SIMNET/CCTT evaluation studies. Kraemer and Rowatt's (1993) *A Review and Annotated Bibliography of Armor Gunnery Training Device Effectiveness Literature* critiques gunnery simulator evaluation practice based on 39 separate evaluations. A paper by Russell (1998) covers the prevalence of the "no significant difference" finding and provides insight into what many evaluators regard as acceptable rules of evidence to make their cases. The reader may act as judge.

Boldovici and Bessemer critiqued both experimental and analytical evaluations of SIMNET. Some of their comments on the experimental evaluations are as follows:

> [The] evaluations incorporated compromises in research design that led to insufficient statistical power, inadequate controls, inappropriate analyses, and irrelevant comparisons.... Inadequate statistical power...was related to the use of too few platoons to detect training effects that may have in fact existed.... The one-shot character of ...evaluations...precluded controlling or randomizing many extraneous variables that could affect evaluation outcomes. (p. 20)

The authors' comments on analytical evaluations are not as critical. They recommend greater use of analytical evaluations. The quote above echoes Boldovici's 1987 book chapter *Measuring Transfer in Military Settings*, in which he summarizes common flaws in training research experiments: not enough subjects, differences between compared groups, different treatments of groups, insufficient amount of practice to affect proficiency, ceiling and floor effects, unreliable test scores, untimely administration of transfer tests, use of inappropriate analyses, and misinterpretation of null results.

Boldovici notes that one of the most common errors is made following a finding of no statistically significant difference (NSD) between groups in multi-group experiments. The null result often occurs in poorly controlled field trials because of inadequate statistical power and is often misinterpreted to mean that the two groups showed equivalent performance. In fact, the null result does not prove equivalence; it does mean that the evaluator cannot say that the scores of the compared groups differ. This subtle difference has tripped up many evaluators. Given a null result, a naïve evaluator may conclude that the alternative forms of training are equally effective and that the logical choice is the least expensive. Do not make this mistake. If it seems improbable that anyone with common sense would make this error, pay close attention to Russell (1998), below.

Kraemer and Rowatt (1993) reviewed 39 studies relating to 15 tank gunnery training devices. One of their goals was to provide sufficient detail to introduce readers to the area rather than simply to identify studies. Findings were broken down in terms of skill acquisition, skill retention, performance prediction, and transfer. Each study was painstakingly reviewed in terms of eight common methodological limitations. These reviews are far more detailed than is common in studies of this ilk.

Table 5-1 summarizes the relative number and percent of the eight types of limitations for the 35 experimental studies in Kraemer and Rowatt's sample. The limitations are listed in order of frequency of occurrence. The most common limitation was 1, small sample size, which occurred in 40% of the studies. This is often the cause of the NSD finding. The next three limitations (unreliable performance measures, groups treated differently, device system errors) were nearly this common. Subjects not random or matched was documented in more than a fifth of the studies. The distribution of these limitations varied across studies: a few had none, most had one or two, a few had more than two. The Overall row indicates that, with a total of 59 limitations distributed across 35 studies, the "average" study had 1.69 of these limitations. The limitations:

1. Small sample size: Small samples result in low statistical power that makes it more difficult to detect true differences between groups. The differences may in fact be real, but statistical tests will not detect them.
2. Unreliable performance measures: Unreliable performance measures do not provide consistent indications of performance and cannot be used to make comparisons between groups.
3. Groups treated differently: If groups participating in an experiment are treated differently (other than for experimental/control treatments), the differential treatment may influence their performance, confounding with the experimental/control treatments.
4. Device system errors: These errors may have a negative effect on subject performance.
5. Subjects not random or matched: Subjects should be randomly assigned or matched prior to an experiment to assure that any differences found between them later can be attributed to the treatment and not to pre-existing differences.
6. Ceiling effect: This generally occurs when the experimental task is too easy. If subjects perform at high levels of proficiency on a task, their scores may show little or no difference.

7. Insufficient amounts of practice: Subjects who are not given sufficient time to practice with an unfamiliar device will still be learning when the experiment takes place and their performance will not reflect the true potential of the device.
8. Floor effect: This generally occurs when the experimental task is too difficult. The inverse of the ceiling effect; if subjects perform at low levels, differences may be undetectable.

Table 5-1. Potential Limitations of Reported Findings on Training Device Effectiveness (Based on N=35 evaluations) (adapted from Kraemer & Rowatt, 1993)

LIMITATION	DESCRIPTION	N	PERCENT
1	Small sample size	14	40
2	Unreliable performance measures	11	31
3	Groups treated differently	10	29
4	Device system errors	10	29
5	Subjects nor random or matched	8	23
6	Ceiling effect	3	9
7	Insufficient amounts of practice	2	6
8	Floor effect	1	3
Overall		59	169

Kraemer and Rowatt explicated their findings in terms of each of the limitations. Their discussion reveals what impact the limitations have on the statistics involved and also suggests actions to prevent the limitations in the first place. For example, regarding limitation 1, it is suggested that evaluators use power analysis to compute sufficient sample sizes to detect effects of a desired magnitude. Refer to the source for a detailed discussion of these limitations and what can be done about them. Some of these issues are covered in the procedural guidance identified in Chapter 6.

Russell (1998) *The "No Significant Difference" Phenomenon as reported in 248 Research Reports, Summaries, and Papers* (fourth edition) is an Internet[69] summary of publications whose main finding was one of no statistically significant difference. Russell includes studies from 1928 to present day, covering mainly educational media. He remarks that his effort is:

> Dedicated with appreciation to all who have submitted works for inclusion in this and past editions of this paper, and also to those who will submit...works for the next (fifth) edition. While this documentation speaks volumes about the futility of these studies, it also acknowledges the fact that the questions about the comparative impacts of the technologies remains of paramount importance. (p. 1)

[69] Russell's summary could be downloaded at http://teleeducation.nb.ca/phenom as this manual went to press.

The point appears to be that all of these media studies have shown NSD—which is taken as evidence that media do not matter. The reader may judge whether or not there is another possible explanation for the NSD finding in some of these studies.

6 PROCEDURAL GUIDANCE

This chapter identifies and summarizes published training effectiveness evaluation guidance from a variety of sources. It is intended to inform the reader about what guidance is available, subjects covered, and possible relevance in evaluating new LSTS. This chapter does not summarize the guidance in enough detail to serve as a substitute for the original source material. Consider this chapter an index to the most important guidance published in the last two decades or so. Use it to survey what is available and select tools that will be helpful in solving new problems. Go to the original sources for details.

The chapter is organized in three sections: evaluation methods, system and program evaluation frameworks, and collective and team training.

Reference List 6-1 (Procedural Guidance) at the end of this chapter contains complete citations for publications cited in this chapter.

Evaluation Methods

This section breaks down evaluation methods using the four-category taxonomy presented in Chapter 3 (experiment, analysis, judgment, and survey). As this manual was written for military training evaluators, the emphasis is on applied rather than academic guidance. It is assumed that readers want practical how-to knowledge. Academic guidance is generally context free and written for general consumption; a good example is Campbell and Stanley (1966). Applied guidance tells evaluators how to solve practical military training evaluation problems; a good example is the Klein, Johns, Perez, and Mirabella (1985) guidebook on comparison-based prediction. Evaluation studies themselves may contain methodological guidance. In a few rare cases, the methodological descriptions are good enough to follow prescriptively. Well-executed studies stand as models. This chapter cites a few examples.

Experiment

Academic

Three key academic works are Campbell and Stanley's classic on quasi-experimental design, Cohen on statistical power analysis, and Cohen and Cohen on multiple regression/correlation. Campbell and Stanley is readily accessible to anyone with a basic background in statistics. The Cohen work on power analysis demands more. The Cohen work on regression/correlation is intended for readers with strong backgrounds in statistics and experimental design.

Cook and Campbell (1979) (*Quasi-Experimentation: Design and Analysis Issues for Field Settings*).[70] This book provides the lexicon and standards commonly accepted among field training evaluators today. It describes and explicates four types of validity: (1) statistical conclusion (2) internal, (3) construct, and (4) external. It describes factors commonly jeopardizing internal validity (history, maturation, testing, instrumentation, statistical regression, differential selection, experimental mortality, selection-maturation interaction) and external validity (reactive effect of testing, interaction effects of selection biases, multiple-treatment interference). It describes and reviews the merits and limitations of experimental designs in terms of validity. It addresses practical problems besetting field experiments and ways to overcome them. It identifies several past examples of true experiments implemented in field settings.

Cohen's 1988 book (*Statistical Power Analysis For the Behavioral Sciences*) describes how to conduct power analysis in hypothesis testing. It provides the rationale underlying power analysis and examples of its application. Contents include concepts of analysis, t test, significance of product moment r, differences between correlation coefficients, test that a proportion is .50 and the sign test, differences between proportions, chi-square, and F tests on means in ANOVA and ANCOVA. This book was followed by a 1992 journal article (*A Power Primer*) evidently written with the evaluation practitioner rather than the theorist in mind. Written in a straightforward style, it is a concise how-to guide for conducting power analyses that includes examples.

Cohen and Cohen (1975) (*Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*) provides background and rationale for multiple regression/correlation analysis, and describes bivariate correlation and regression, multiple regression and correlation, analysis of covariance, other multivariate methods, sets of independent variables, nominal or qualitative scales, quantitative scales, and issues of missing data, interactions, and repeated measurement.

Applied

Pfeiffer and Browning (1984) and Morrison and Hoffman (1992) have published widely-cited applied works on experimental design, both with an emphasis on transfer designs.[71] Boldovici and Bessemer (1999) offer a set of rules for conducting valid experiment- and judgment-based evaluations. Kass presents guidelines and a job aid to help evaluators design and conduct valid field experiments.

Pfeiffer and Browning (1984) (*Field Evaluations of Aviation Trainers*) provide an excellent overview of several alternative experimental designs and methods that have been used for aviation TEA; for

[70] See also Campbell and Stanley (1966) *(Experimental and Quasi-Experimental Designs for Research)* and Cook and Campbell (1976) *(The Design and Conduct of Quasi-Experiments and True Experiments in Field Settings)*. The 1966 work (an 84-page book) was first published in 1963 in N.L. Gage (Ed.) *Handbook of Research on Teaching*. From the earlier to the later publications, each of these is a successively more elaborate treatment of ideas presented in the earlier works.

[71] Pfeiffer and Browning's study was written for the aviation community and Morrison and Hoffman's for the tank gunnery community. As noted earlier in this manual, these two training communitites make the greatest use of transfer experiments.

example, transfer experiments, quasi-experiments, and analytic studies. Selecting a proper research design depends upon the purpose of the evaluation and training constraints. Despite its title, the guidance applies beyond aviation training evaluation. They cover issues affecting selection of designs and dependent measures, and obstacles to conducting good field evaluations. They present examples of experimental, quasi-experimental, and analytical designs for various training situations and provide guidance to match evaluation designs to field situations.

Morrison and Hoffman (1992) (*A User's Introduction To Determining Cost-Effective Tradeoffs among Tank Gunnery Training Methods*) discuss how to make tradeoffs among relative amounts of training on alternative (2 or more) devices and actual equipment. The authors describe several transfer of training experimental designs (2-group, multi-group, groups-by trials, multidimensional). The report describes the support requirements to obtain data using the methods specified. The authors emphasize the use of performance-based data, but also describe an alternative judgment-based method (simulated transfer) to generate surrogate performance data. As previously, this report applies beyond the subject of its title.[72]

Boldovici and Bessemer (1999) (*The Elements of Training Evaluation*) offer evaluation guidance in the form of 15 declarative rules that apply to experimental and judgment-based evaluation methods.[73] Other topics covered are ratings, how to deal with null results, and increasing statistical power. The guidance consolidates the many methodological ideas and critiques these two authors have contributed to the evaluation literature over the years. Some of these ideas have already been noted. Among them is an emphasis on the importance of the psychometric quality (validity and reliability) of data. The authors give this greater weight than the method of obtaining data. Frequent critics of field experiments, they argue that judgment-based and analytical evaluation can provide better results. Their evaluation guide tells how to make the most of such data.

Kass (1997, June/July) (*Design of Valid Operational Tests*[74]) presents a framework to organize and relate good test practices to maximize test and experiment design validity. The paper offers a definition of validity, identifies 19 threats to it, and discusses how to design tests to maximize validity. The analysis is based on Cook and Campbell (1979). The 19 possible threats to validity are based on the combination of experimental components (treatments, test units, effects, trials, analyses) and design validity (statistical validity, single group, multiple groups, operational validity). Threats are: violating assumptions of statistical tests; error rate problems; low power statistical analysis; variability in system, player unit, data collection, or trial conditions; changes over time to treatments,

[72] Both Pfeiffer and Browning and Morrison and Hoffman emphasize the experimental method, but both also endorse judgment-based methods under certain circumstances. Morrison and Hoffman develop their ideas on simulated transfer in this report. Pfeiffer and Horey (1988) describe four classes of judgment-based methods for forecasting and evaluating training effectiveness; this report is described later in this chapter.

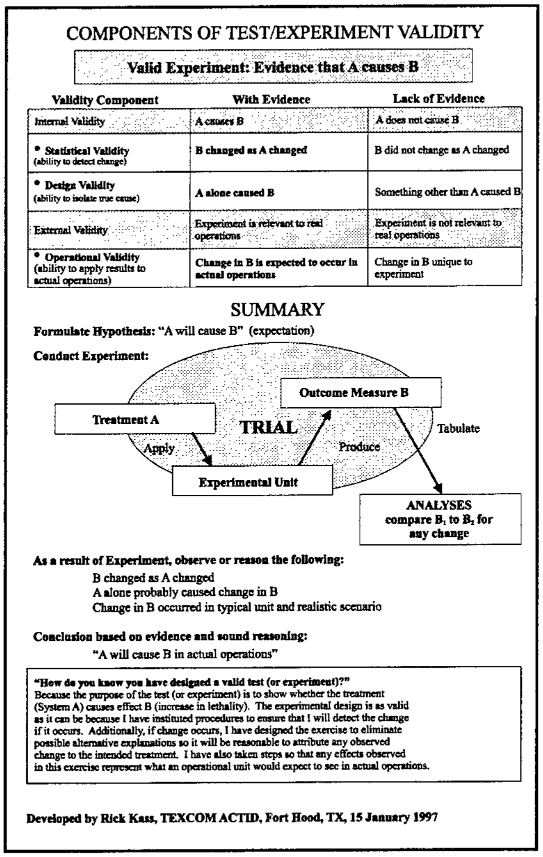
[73] At the time this manual went to press, Boldovici and Bessemer's rules were in draft form. They may change. The rules, elaborated in separate paragraphs, are: (1) Consider testing the alternative to the null hypothesis; (2) Specify the risk the evaluation customer is willing to take of erroneously detecting no differences between the compared groups' scores; (3) Perform power analyses to determine the number of observations necessary to detect differences between the scores of compared groups; (4) Assign soldiers or units randomly to the compared kinds of training (treatments); (5) Establish that the compared groups do not differ significantly in ways that might affect outcomes; (6) Treat the compared groups identically during the evaluation in all respects save treatments; (7) The reliability of the posttests, that is, the tests administered after training the compared groups, must be at least 75%; (8) The difficulty of the posttests must permit few and preferably no scores greater than 75% or less than 25%; (9) Allow some time to pass between the end of training and the beginning of testing; (10) Administer more than one posttest; (11) The time between the end of training and the beginning of testing must be identical for the compared groups; (12) Use conventional analyses of raw scores to estimate training effects; (13) Perform separate analyses of training-sensitive and training-insensitive test items; (14) Interpret null results in terms of confidence intervals; (15) Report generalizability estimates.

[74] Kass, a member of the military testing community, refers to the studies he describes as "tests" rather than experiments. Within the lexicon used in the present report, a test is considered to be a type of experiment and hence Kass' ideas generalize to experiments as well as tests.

player units, data collection, or trial conditions; differences in player units, data collection, or trial conditions; and nonrepresentative system, units, measures, scenarios, or sites. The validity framework can be used as a checklist when designing test plans and experiments, to compare alternate test designs, and for training data collectors and test player units.

Kass (1997) in *Test Officer's Guide for Designing Valid Tests and Experiments* provides a compact job aid to help evaluators identify and deal with the 19 threats. The job aid consists of a single 8-1/2 X 11" sheet printed on both sides and folded down the middle so that it fits into a large pocket. This makes application of the ideas in the analysis fairly straightforward. The job aid is reproduced in Figure 6-1.

Figure 6-1. Test Officer's Guide for Designing Valid Tests and Experiments (Front)



COMPONENTS OF TEST/EXPERIMENT VALIDITY

Valid Experiment: Evidence that A causes B

Validity Component	With Evidence	Lack of Evidence
Internal Validity	A causes B	A does not cause B
• **Statistical Validity** (ability to detect change)	**B changed as A changed**	B did not change as A changed
• **Design Validity** (ability to isolate true cause)	**A alone caused B**	Something other than A caused B
External Validity	Experiment is relevant to real operations	Experiment is not relevant to real operations
• **Operational Validity** (ability to apply results to actual operations)	**Change in B is expected to occur in actual operations**	Change in B unique to experiment

SUMMARY

Formulate Hypothesis: "A will cause B" (expectation)

Conduct Experiment:

As a result of Experiment, observe or reason the following:

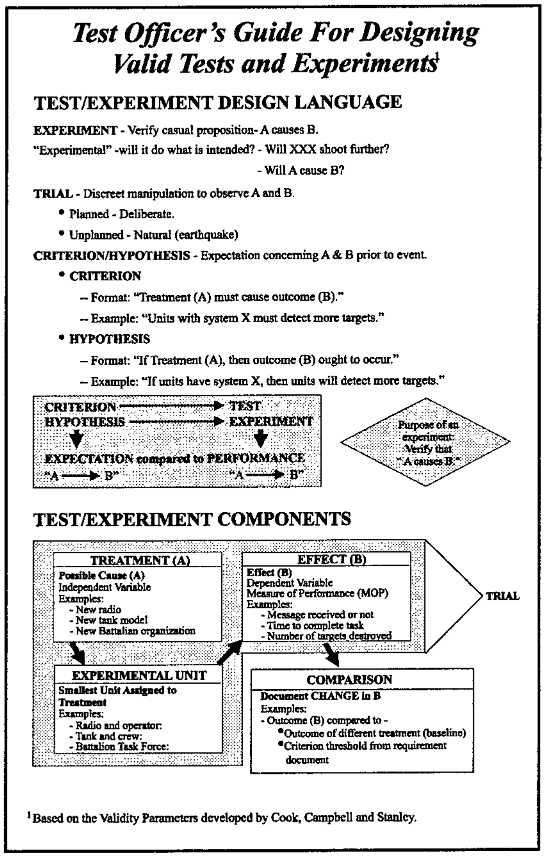
B changed as A changed
A alone probably caused change in B
Change in B occurred in typical unit and realistic scenario

Conclusion based on evidence and sound reasoning:

"A will cause B in actual operations"

"How do you know you have designed a valid test (or experiment)?"
Because the purpose of the test (or experiment) is to show whether the treatment (System A) causes effect B (increase in lethality). The experimental design is as valid as it can be because I have instituted procedures to ensure that I will detect the change if it occurs. Additionally, if change occurs, I have designed the exercise to eliminate possible alternative explanations so it will be reasonable to attribute any observed change to the intended treatment. I have also taken steps so that any effects observed in this exercise represent what an operational unit would expect to see in actual operations.

Developed by Rick Kass, TEXCOM ACTID, Fort Hood, TX, 15 January 1997



Test Officer's Guide For Designing Valid Tests and Experiments[1]

TEST/EXPERIMENT DESIGN LANGUAGE

EXPERIMENT - Verify casual proposition- A causes B.

"Experimental" -will it do what is intended? - Will XXX shoot further?
- Will A cause B?

TRIAL - Discreet manipulation to observe A and B.

- Planned - Deliberate.
- Unplanned - Natural (earthquake)

CRITERION/HYPOTHESIS - Expectation concerning A & B prior to event.

- **CRITERION**
  - -- Format: "Treatment (A) must cause outcome (B)."
  - -- Example: "Units with system X must detect more targets."
- **HYPOTHESIS**
  - -- Format: "If Treatment (A), then outcome (B) ought to occur."
  - -- Example: "If units have system X, then units will detect more targets."

CRITERION → TEST
HYPOTHESIS → EXPERIMENT
EXPECTATION compared to PERFORMANCE
"A → B" "A → B"
Purpose of an experiment: Verify that "A causes B."

TEST/EXPERIMENT COMPONENTS

TREATMENT (A)
Possible Cause (A)
Independent Variable
Examples:
- New radio
- New tank model
- New Battalian organization

EFFECT (B)
Effect (B)
Dependent Variable
Measure of Performance (MOP)
Examples:
- Message received or not
- Time to complete task
- Number of targets destroyed

EXPERIMENTAL UNIT
Smallest Unit Assigned to Treatment
Examples:
- Radio and operator:
- Tank and crew:
- Battalion Task Force:

COMPARISON
Document CHANGE in B
Examples:
- Outcome (B) compared to -
  - Outcome of different treatment (baseline)
  - Criterion threshold from requirement document

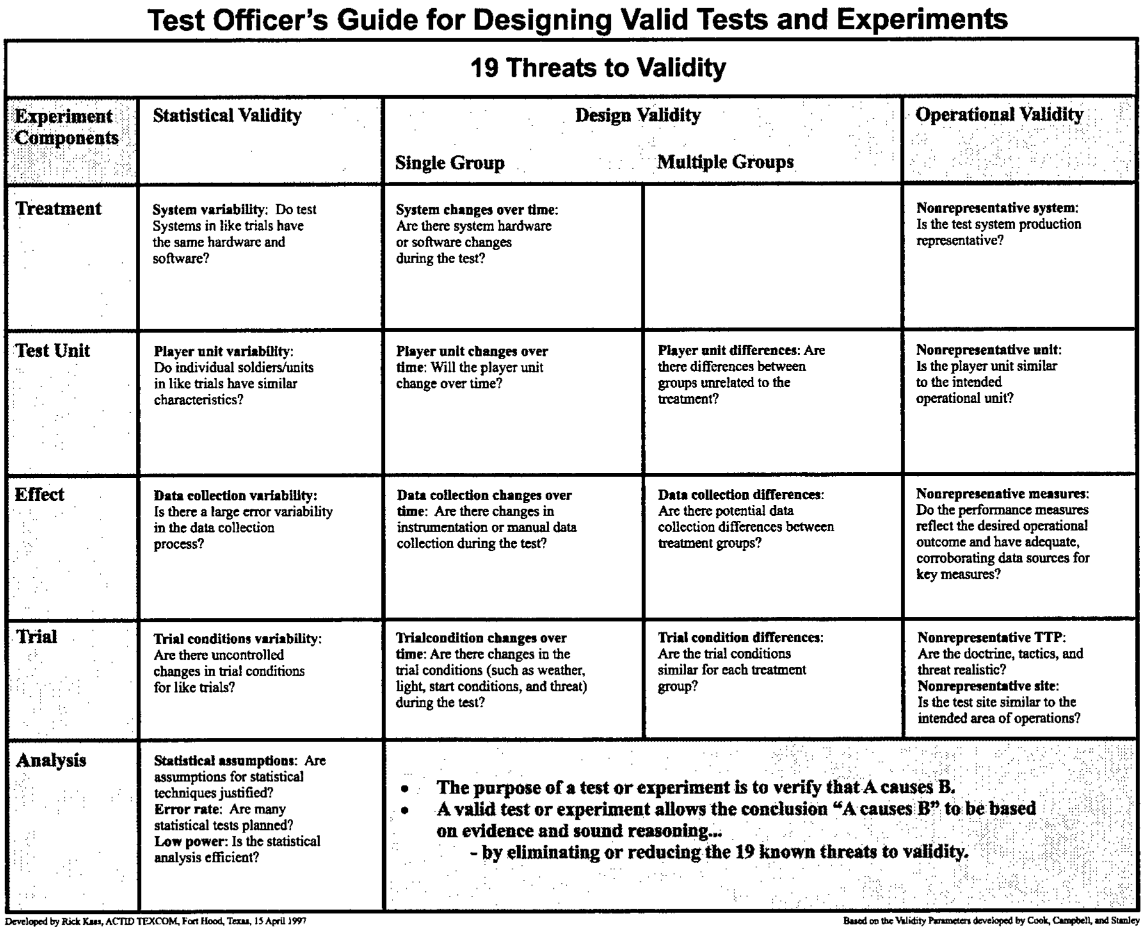
[1] Based on the Validity Parameters developed by Cook, Campbell and Stanley.

Analytical Evaluation

Chapter 3 noted that there is no universal definition of analytical evaluation. It then described three ways that analysis was commonly used (Evaluate, Compare, Optimize), several different evaluation strategies, and some formal analytical evaluation methods. This section focuses on the formal analytical evaluation methods. The published guidance for formal analytical methods is large and confusing. Hundreds of studies have been published on the subject in the last two decades. The majority of these are applied studies conducted under military contract. A 1994 review by Muckler and Finley sorts out this literature and is recommended to those interested in the field.[75] This section begins, on a smaller scale, with Pfeiffer and Horey (1988), and discusses several of the more prominent methods in enough detail that readers should be able to estimate their utility. Guidance for a few other methods is then described. This section focuses on a small fraction of the

[75] Muckler and Finley (1994a,b) is a two-volume review that describes and compares the most significant of these methods clearly and concisely from a historical perspective for the decade 1970-1990; this review is recommended to readers interested in the methods and their historical development. Volume I (Muckler & Finley, 1994a) contains a literature review and analysis and volume II (1994b) contains a 175-item annotated bibliography that covers the essential literature in the field.

Figure 6-1. Test Officer's Guide for Designing Valid Tests and Experiments (Back)



Test Officer's Guide for Designing Valid Tests and Experiments

19 Threats to Validity

Experiment Components	**Statistical Validity**	**Design Validity**		**Operational Validity**
		Single Group	**Multiple Groups**	
Treatment	**System variability:** Do test Systems in like trials have the same hardware and software?	**System changes over time:** Are there system hardware or software changes during the test?		**Nonrepresentative system:** Is the test system production representative?
Test Unit	**Player unit variability:** Do individual soldiers/units in like trials have similar characteristics?	**Player unit changes over time**: Will the player unit change over time?	**Player unit differences**: Are there differences between groups unrelated to the treatment?	**Nonrepresentative unit:** Is the player unit similar to the intended operational unit?
Effect	**Data collection variability:** Is there a large error variability in the data collection process?	**Data collection changes over time**: Are there changes in instrumentation or manual data collection during the test?	**Data collection differences**: Are there potential data collection differences between treatment groups?	**Nonrepresenative measures**: Do the performance measures reflect the desired operational outcome and have adequate, corroborating data sources for key measures?
Trial	**Trial conditions variability:** Are there uncontrolled changes in trial conditions for like trials?	**Trialcondition changes over time**: Are there changes in the trial conditions (such as weather, light, start conditions, and threat) during the test?	**Trial condition differences**: Are the trial conditions similar for each treatment group?	**Nonrepresentative TTP**: Are the doctrine, tactics, and threat realistic? **Nonrepresentative site**: Is the test site similar to the intended area of operations?
Analysis	**Statistical assumptions**: Are assumptions for statistical techniques justified? **Error rate**: Are many statistical tests planned? **Low power**: Is the statistical analysis efficient?	• **The purpose of a test or experiment is to verify that A causes B.** • **A valid test or experiment allows the conclusion "A causes B" to be based on evidence and sound reasoning...** **- by eliminating or reducing the 19 known threats to validity.**		

Developed by Rick Kass, ACTID TEXCOM, Fort Hood, Texas, 15 April 1997

Based on the Validity Parameters developed by Cook, Campbell, and Stanley formal methods. The selection was based on the method's relative simplicity, adequacy of documentation, and apparent use. The methods described in these documents are FORTE, Conjoint Analysis, DEFT (Device Effectiveness Forecasting Technique), Simulated Transfer, Comparison-Based Prediction, and the Training Mix Model.

Pfeiffer and Horey (1988) (*Analytic Approaches To Forecasting and Evaluating Training Effectiveness*) describes, compares, and contrasts four classes of methods for forecasting and evaluating training effectiveness:

- Index techniques (checklist, display evaluation index, analytic profile system, instructional quality inventory)
- Magnitude techniques (simulated transfer, FORTE, conjoint analysis, DEFT)
- Proximity techniques (simulated training capability, task commonality analysis, fidelity analysis, device handling qualities, multitrait-multimethod matrix, comparison-based prediction)
- Interlocking techniques (multiattribute utility analysis, multidimensional scaling analysis, training interlock measure, system operability measurement algorithm).

The authors make the point that these methods can be of value during the device acquisition process, when opportunities to conduct experimental research and evaluation are limited.

Klein, Johns, Perez, and Mirabella (1985) (*Comparison-Based Prediction of Cost and Effectiveness of Training Devices: A Guidebook*) is a how-to guide for applying the comparison-based prediction method to predict the cost and training effectiveness of new systems. The method extrapolates the cost and training effectiveness of the new system based on its similarities to and differences from an existing system. The procedure is similar to the comparative market approach used in real estate appraisal.

Djang, Butler, Laferriere, and Hughes (1993) (*Training Mix Model*) describe an analytical method to optimize the mix of field training and training using training aids, devices, simulators, and simulations in terms of cost-effectiveness. The "training mix model" is a computer program that incorporates the expected cost of acquiring and using training systems with their expected effectiveness in terms of ability to train required tasks. TRAC-WSMR continues to develop, apply, and refine this method.

Judgment and Survey

Judgment-based evaluations and Surveys gather data in the same ways; that is, with questionnaire, interview, and observation. Most of the guidance that applies to one applies to the other. The main difference between them is scale. They are treated together here. Some of the guidance on the conduct of large-scale surveys is inapplicable to smaller, judgment-based evaluations. Use common sense to decide which of the guidance applies to each evaluation method.

The published guidance for these methods is good. The scope of individual publications ranges from big picture (e.g., design and conduct of large-scale field surveys) to small (e.g., how to design rating scales). Fowler (1993) has published a straightforward guide on survey research methods. Litwin (1995) addresses how to measure survey reliability and validity. Bouchard (1976) published a widely-used work on field research methods, some of which are commonly used in judgment-based evaluations and surveys. Patton (1987) covers similar ground. Other guidance covers the design of questionnaires (Babbit and Nystrom, 1989a,b), measurement of attitudes (Henerson, Morris, and Fitz-Gibbon, 1987), and construction of rating scales (Spector, 1992).

Fowler (1993) (*Survey Research Methods*) is a how-to guide for conducting surveys. Its coverage includes sampling, dealing with non-responses, methods of data collection, designing and evaluating survey questions, interviewing, data analysis, survey error, and ethical issues. The book provides standards and practical procedures for surveys designed to provide statistical descriptions of people by asking questions, usually of a sample. The book explains how each aspect of a survey can affect its precision, accuracy, and credibility.

Litwin (1995) (*How To Measure Survey Reliability and Validity*) is a concise how-to guide for creating valid and reliable surveys. It includes an overview of psychometrics, reliability (test-retest, alternate-form, internal consistency, inter-observer), validity concepts (face, content, criterion, construct), scaling and scoring, use of code books, pilot testing, and multicultural issues.

Bouchard (1976) (*Field Research Methods: Interviewing, Questionnaires, Participant Observation, Systematic Observation, Unobtrusive Measures*) deals with the larger question of field research, rather than with surveys or judgment. His discussion of these methods is pertinent. Also, he deals with the special characteristics of field settings (intensity, range, frequency and duration, natural time constant, natural units, setting effects) and their special difficulties. (Some of these issues were raised in Chapter 5.) Bouchard provides descriptions of and guidelines for applying each of the field research methods: interviewing, questionnaires, participant observation, systematic observation, and unobtrusive measures.

Patton (1987) (*How To Use Qualitative Methods in Evaluation*) is a straightforward, non-theoretical how-to guide for the use of qualitative methods (open-ended interviews, direct observation, written documents). Its coverage includes when to use qualitative methods, designing qualitative evaluations, fieldwork and observation, depth interviewing, analyzing and interpreting data, and making methods decisions.

Babbitt and Nystrom (1989a,b) (*Questionnaire Construction Manual* and *Annex*) is a two-volume guide on how to create questionnaires. The Annex is a literature survey and bibliography on questionnaire construction. The manual describes current methods based on research for developing questionnaires. The manual was designed to guide individuals who develop and/or administer questionnaires as part of Army field tests and evaluations but its content applies to many nonmilitary applications. Key concepts covered are questionnaire construction, questionnaire administration, attitude scales, scaling techniques, response anchoring, response alternatives, pretesting questionnaires, survey interviews, demographic characteristics, continuous and circular scales, questionnaire layout, branching, scale points, response alternatives, and item wording.

Henerson, Morris, and Fitz-Gibbon (1987) (*How To Measure Attitudes*) is a straightforward how-to guide for measuring attitudes. It covers measurement of attitudes and attitude change, essential preliminary questions, collecting attitude information, finding an existing measure, developing measures, attitude rating scales, interviews, written reports, observation procedures, sociometric instruments, validity and reliability of attitude instruments, displaying data.

Spector (1992) (*Summated Rating Scale Construction*) is a how-to guide for writing items and creating valid and reliable rating scales. It covers theory of summated rating scales, defining the construct, designing the scale, item analysis, validation, reliability, and norms.

Boldovici and Bessemer (1999) (*The Elements of Training Evaluation*), cited earlier, includes a section on the use of ratings during evaluation.

System and Program Evaluation Frameworks

Chapter 2 began building an evaluation framework for LSTS by asking basic questions about the why, who, what, where, how, and when of evaluation. An evaluation framework offers answers to these questions and expresses an evaluation philosophy. As this manual was being written there was no such framework for evaluating LSTS. To build the new framework, the author reviewed analogous frameworks in human factors, public education, and military training. Certain elements of these frameworks were adopted in the framework described in Chapter 8. These frameworks remain of interest in their own right. By reviewing them, the reader can see how they influenced the framework presented in this manual and modify or customize that framework to suit particular circumstances.

Of the many books written on human factors evaluation during system development, Meister and Rabideau (1965) and Meister (1986) have remained oft-cited classics that reflect the human factors professional's point of view. Kirkpatrick (1976) is widely regarded as the standard work on training program evaluation. Herman, Morris, and Fitz-Gibbon (1987) wrote a how-to guide on this subject for public education. Guides by Semple (1974) and Hall, Rankin, and Aagard (1976) were developed expressly for military training effectiveness evaluation.

Human factors evaluations are analogous to training system evaluations because they require system developers to acknowledge human needs in terms of the human factors aspects of a design or its training effectiveness as expressed in the design. Both human factors and training communities often find that their attempts to influence designs compete with and may conflict with hardware and software development. Hence, the strategies proposed by human factors professionals may serve as models for use by training developers and evaluators. Meister and Rabideau (1965) (*Human Factors Evaluation in System Development*) is a guide for conducting human factors evaluations in the field or operational setting. It describes several human factors analysis and evaluation methods; for example, functional analysis, human engineering evaluation (examination of design criteria, drawings, diagrams, operator and group procedures, mockups, developmental tests), system performance evaluation (simulation and operational testing, R&D testing, field testing). It tells how to plan a performance evaluation and sketches data collection methods (method selection, direct methods, indirect methods), data analysis, and evaluation of production. Much of the methodology can be extrapolated to training effectiveness evaluation; for example, Chapters 3 (functional analysis) and Chapters 7 and 8 (data collection methods).

Meister (1986) (*Human Factors Testing and Evaluation*) covers testing during system development, laboratory research versus performance testing, use of mockups, developmental and operational testing, test plans measurement methods (job performance observation, self-report, interview, questionnaire, ratings, subjective methods, activity analysis, objective measures), environmental testing, special measurement methods (human error, computerized systems and software, maintenance performance, team performance, workload evaluation, training systems and devices, transfer), testing literature, test planning, measurement models, training effectiveness evaluation, human engineering reviews, maintainability. Much of the methodology described can be extrapolated to training effectiveness evaluation; for example, Chapters 4 (measurement methods), 6 (measurement problems).

Kirkpatrick (1976) (*Evaluation of Training*) is discussed in Chapter 7. It presents a strategy for evaluating training programs based upon four variables: (1) reaction (how well did participants like program?), (2) learning (what principles, facts, and techniques did students learn?), (3) behavior (what changes in job behavior resulted from the program?), (4) results (what were the tangible results of the program in terms of reduced cost, improved quality, improved quantity, etc.?). It includes many examples of data collection protocols.

Herman, Morris, and Fitz-Gibbon (1987) (*Evaluator's Handbook*) is a practical, how-to guide for planning and conducting formative and summative program evaluations. Contents: (1) establishing parameters of an evaluation (evaluation framework, determining approach, what to measure or observe), (2) formative or summative evaluation (set boundaries, select methods, collect and analyze information, report), (3) guide to conduct formative evaluation, (4) guide to conduct summative evaluation, (5) guide to conduct small experiment.

Semple (1974) (*Guidelines for Implementing Training Effectiveness Evaluations*) provides a framework for conducting training evaluations of training programs, devices, and systems. It is broad and generalizable. It describes the training evolution process—and what kind of information can be obtained as training undergoes development. It identifies a set of common baseline assumptions about evaluation (e.g., they are experimentally-based, use the transfer paradigm, etc.) and encourages reality testing before assuming them true. Four phases of evaluation are described: planning, execution, analysis, and documentation. It describes Jeantheau's four levels of evaluation (qualitative, non-comparative, comparative, transfer), with each successive level providing stronger evidence. It suggests key issues to address at each phase of evaluation and discusses methods to employ based on Jeantheau's framework.

Hall, Rankin, and Aagard (1976) (*Training Effectiveness Assessment, Volume II: Problems, Concepts and Evaluation Alternatives*) presents an analysis and description of problems inherent in training evaluation in terms of external factors (attitudes toward evaluation, administration, personnel). It describes process and product evaluation, choosing measures and obtaining evaluation data, designing evaluation plans, data quality (reliability, validity), data gathering options and procedures (tests, questionnaires, interviews, records), and interpretation of data.

Collective and Team Training

Large-scale training simulations such as the CCTT are used primarily for collective training—to train groups of individuals (e.g., crews, teams, units) who must work together and coordinate their activities. Collectives often comprise hundreds or thousands of people. Team training is a type of collective training involving small groups; for example, fewer than a dozen people. A team may be an aircrew, an armored vehicle crew, or a group of senior leaders who must coordinate their activities to wage a battle. Large-scale training simulations used to train senior leaders (such as the JSIMS) are concerned with team training. To evaluate the effectiveness of LSTS, the evaluator must find ways to evaluate collective training, team training, or both. The methods for doing this are still immature. However, in recent years a number of documents have been published that offer guidance and evaluation tools.

Collective Training

In 1994, the Army Research Institute published a book-length report (Holz, Hiller, and McFann [Eds.]: *Determinants of Effective Unit Performance*) that deals with the assessment of unit training readiness. While this entire report is of interest, two chapters in Section 1 (*Measuring Unit Performance*) are noteworthy. Lewman, Mullen, and Root (*A Conceptual Framework for Measuring Unit Performance*) describe a framework for evaluating unit performance at the NTC based on unit missions, echelons, and critical tasks, using analysis and SME judgments. Task structure, standards, conditions, and measurement protocols are described. This approach could be generalized to the measurement of unit combat performance in other settings. Fober, Dyer, and Salter (*Measurement of Performance at the Joint Readiness Training Center [JRTC]: Tools of Assessment*) describe performance measures used at JRTC: (1) training and evaluation outlines, (2) take-home packages, and (3) AARs. Measures (1) and (2) are prepared by O/Cs and AARs are conducted by JRTC and videotaped; all of this material may be provided to personnel participating in JRTC exercises. The chapter also discusses measurement methods and data collection instruments (e.g., rating scales, checklists, formats, and methods used in conducting AARs.

Turnage, Houser, and Hofmann (1990) (*Assessment of Performance Measurement Methodologies for Collective Military Training*) is a wide-ranging review of the state of the art in military collective performance measurement methodologies (e.g., ARTEP, AAR, SIMNET) and their strengths and weaknesses. It includes a review of team and collective training research and describes the methods used to conduct collective training in the Army (e.g., MILES, CATTS, SIMCAT, etc.). This source shows how the Army has traditionally assessed collective training.

Some publications cited in Chapter 4 for the SIMNET/CCTT case study may also be of value for collective training assessment. See Meliza and Tan (1992); Meliza, Bessemer, and Tan (1992); Meliza, Bessemer, Burnside, and Shlechter (1992); and Meliza (1993).

Team Training

The TARGETS and TOMs team training assessment methods were discussed in Chapter 4. For TARGETs, see Fowlkes, Lane, Salas, Franz, and Oser, R. (1994) and Dwyer, Fowlkes, Oser, Salas, and Lane (1997). For TOMs, see Dwyer, Fowlkes, Oser, and Salas (1996).

Cannon-Bowers and Salas (1997) (*A Framework for Developing Team Performance Measures in Training*) describes a process for measuring team performance. The paper discusses the nature of team training and its measurement, means of data collection (observational checklist, computer input, automatic data recording), and principles for creating measures of effectiveness. The authors contend that measures used should reflect multiple levels (e.g., individual, team); assess both process and outcomes; describe, evaluate, and diagnose performance; and provide a basis for remediation. The descriptive framework presents suggested data collection instruments and analytical methods for obtaining process and product measures for individual and team performance.

Garlinger and Fallesen (1988) (*Review of Command Group Training Measurement Methods*) reviews the available performance measurement tools for command group training. Techniques (self-assessment, peer assessment, SME observation, ARTEP, probes, battle outcome data, etc.) are discussed and compared. Suggestions are made for an evaluative strategy.

Reference List 6-1. Procedural Guidance

Babbitt, B.A. & Nystrom, C.O. (1989a). *Questionnaire construction manual.* RP 89-20. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA212365)

Babbitt, B.A. & Nystrom, C.O. (1989b). *Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography.* RP 89-21. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA213255)

Boldovici, J.A. & Bessemer, D.W. (1999). *The elements of training evaluation.* Special Report. Alexandria, VA: U.S. Army Research Institute for Behavioral and Social Sciences.

Bouchard, T.J. (1976). Field research methods: Interviewing, questionnaires, participant observation, systematic observation, unobtrusive measures. In M.D. Dunnette (Ed.) *Handbook of Industrial and Organizational Psychology.* Chicago, IL: Rand McNally.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* In N.L. Gage (Ed.). *Handbook of Research on Teaching.* Chicago, IL: Rand-McNally.

Campbell, D.T. & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally. (Also published as Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental design in teaching.* In N.L. Gage (Ed.). Handbook of Research on Teaching. Chicago, IL: Rand-McNally.)

Cannon-Bowers, J.A. & Salas, E. (1997). A framework for developing team performance measures in training. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team Performance Assessment and Measurement: Theory, Methods and Applications.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cook, T.D. & Campbell, D.T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.). *Handbook of Industrial and Organizational Psychology* (pp. 223-326). Chicago, IL: Rand McNally.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago, IL: Rand McNally.

Djang, P.A., Butler, W.A., Laferriere, R.R., & Hughes, C.R. (1993). *Training mix model.* TRAC-WSMR-TEA-93-035. White Sands Missile Range, NM: TRADOC Analysis Center. (ADB178428)

Dwyer, D.J., Fowlkes, J., Oser, R.L., & Salas, E. (1996). Panel on multi-service distributed training testbed, DIS training of military teams/groups: Case study results using distributed interactive simulation for close air support. *Proceedings of the 1996 International Training Equipment Conference.* The Hague, Netherlands, 371-380.

Dwyer, D.J., Fowlkes, J.E., Oser, R.L., Salas, E., & Lane, N.E. (1997). Team performance measurement in distributed environments: the TARGETs methodology. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team Performance Assessment and Measurement: Theory, Methods and Applications* (pp. 137-153). Hillsdale, NJ: Lawrence Erlbaum.

Fober, G.W., Dyer, J.L., & Salter, M.S. (1994). Measurement of performance at the joint training center: Tools of assessment. In. R.F. Holz, J.H. Hiller, & H.H. McFann (Eds.) (1994). *Determinants of effective unit performance: Research on measuring and managing unit training readiness* (pp. 39-70). Book. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Fowler, F.J. (1993). *Survey research methods.* Beverly Hills, CA: Sage Publications

Fowlkes, J.E., Lane, N.E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology, 6*, 47-61.

Gage, N.L. (Ed.) (1963). *Handbook of research on teaching.* Chicago, IL: Rand-McNally.

Garlinger, D.K. & Fallesen, J.J. (1988). *Review of command group training measurement methods.* TR 798. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA201753)

Hall, E.R, Rankin, W.C., & Aagard, J.A. (1976) *Training effectiveness assessment, volume II: Problems, concepts and evaluation alternatives.* TAEG Report 39. Orlando, FL: Training Analysis and Evaluation Group. (ADA036518)

Henerson, M.E., Morris, L.L., & Fitz-Gibbon, C.T. (1987). *How to measure attitudes.* (Volume 6 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Herman, J.L., Morris, L.L., & Fitz-Gibbon, C.T. (1987). *Evaluator's handbook.* (Volume 1 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Holz, R.F., Hiller, J.H., & McFann, H.H. (Eds.) (1994). *Determinants of effective unit performance: Research on measuring and managing unit training readiness.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Jeantheau, G.G. (1971). *Handbook for training systems evaluation.* NAVEDTRACEN 66-C-0113-2. Darien, CT: Dunlap & Associates. (AD733962)

Kass, R., (1997). *Test officer's guide for designing valid tests and experiments* (job aid). Ft. Hood, TX: TEXCOM Combined Arms Test Center.

Kass, R., (1997, June/July). Design of valid operational tests. *ITEA Journal.*

Kirkpatrick, D.L. (1976). Evaluation of training. In Craig, R.L. (Ed.) (1976). *Training and development handbook: A guide to human resource development* (pp18-1-18-27). New York, NY: McGraw-Hill.

Klein, G.A., Johns, P., Perez, R., & Mirabella, A. (1985). *Comparison-based prediction of cost and effectiveness of training devices: A guidebook.* RP 85-29. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA170941)

Lewman, T., Mullen, W.J., & Root, J. (1994). A conceptual framework for measuring unit performance. In. R.F. Holz, J.H. Hiller, & H.H. McFann (Eds.). *Determinants of effective unit performance: Research on measuring and managing unit training readiness* (pp. 17-38). Book. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Litwin, M.S. (1995). *How to measure survey reliability and validity.* Beverly Hills, CA: Sage Publications.

Meister, D. & Rabideau, G.F. (1965). *Human factors evaluation in system development.* New York: Wiley.

Meister, D. (1986). *Human factors testing and evaluation.* New York: Elsevier.

Meliza, L.L. & Tan, S.C. (1992). *SIMNET unit performance assessment system (UPAS) version 2.5 user's guide.* ARI-RP-96-05. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA318046)

Meliza, L.L. (1993). *Simulation networking/ training requirements relational database: User's guide.* RP 94-01. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275634)

Meliza, L.L., Bessemer, D.W., & Tan, S.C. (1992). *Unit performance assessment systems development.* TR 1008. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285805)

Meliza, L.L., Bessemer, D.W., Burnside, B.L., & Shlechter, T.M. (1992). *Platoon-level after action review aids in the SIMNET unit performance assessment system (UPAS).* TR 956. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA254909)

Morrison, J.E. & Hoffman, R.G. (1992). *A user's introduction to determining tradeoffs among tank gunnery training methods.* ARI RN 92-29. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA250029)

Muckler, F.A. & Finley, D.L. (1994a). *Applying training system estimation models to army training, Volume I: Analysis of the literature.* ARL-TR-463. Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory. (ADA283021)

Muckler, F.A. & Finley, D.L. (1994b). *Applying training system estimation models to army training, Volume II: An annotated bibliography 1970-1990.* ARL-TR-463. Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory. (ADA283022)

Patton, M.Q. (1987). *How to use qualitative methods in evaluation.* (Volume 4 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Pfeiffer, M.G. & Browning, R.F. (1984). *Field evaluations of aviation trainers.* NTSC TR-157. Orlando, FL: Naval Training Systems Center. (ADB083584).

Pfeiffer, M.G. & Horey, J.D. (1988). *Analytic approaches to forecasting and evaluating training effectiveness.* NTSC TR-88-027. Orlando, FL: Naval Training Systems Center. (ADB129158)

Semple, C.A. (1974). *Guidelines for implementing training effectiveness evaluations.* NTEC TR NAVTRAEQUIPCEN 72-C-0209-3. Orlando, FL: Naval Training Equipment Center.

Spector, P.E. (1992). *Summated rating scale construction.* Beverly Hills, CA: Sage Publications.

Turnage, J.J., Houser, T.L., & Hofmann, D.A. (1990). *Assessment of performance measurement methodologies for collective military training.* RN 90-126. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA227971)

7 EVALUATION CRITERIA

Evaluation criteria are the measures collected during an evaluation whose values are used to decide the outcome of the evaluation. Dependent variables in experimental research are one type of evaluation criteria. The chapter begins by expanding the definition of evaluation criteria beyond dependent variables used with the experimental method into the domain of M&S (Modeling and Simulation). It discusses how evaluation criteria vary with evaluation method and with small- versus large-scale evaluations. It discusses how the perspectives of training evaluators, program managers, and the M&S community differ in terms of evaluation criteria, and the need to coordinate among these parties during training system development. The final section develops a set of evaluation criteria for use in evaluating the training effectiveness of an LSTS.

Defining Evaluation Criteria

Evaluation criteria include but go beyond the dependent variables associated with the experimental method. Moreover, they vary with evaluation scale (small or large) and perspective; that is, who is doing the evaluation and for what purpose.

Criteria and Evaluation Methods

The SIMNET/CCTT evaluations in TCEF are summarized in Table 7-1 by evaluation method, author, and dependent variables used. These evaluations were described in Chapter 4 and this table is based on Table 4-1 in Chapter 4. Within the Method column, evaluations are listed in order of year of publication.

Many of the entries appearing in the Dependent Variables column were discussed in Chapter 4 so they will not be discussed here. However, a few things are worth mentioning. First, note that the table includes evaluations for four different evaluation methods (experiment, analysis, judgment, and survey) and that there are entries in the Dependent Variables column for all of them. Dependent variables are usually associated with the experimental method. However, it is obvious from the entries in this table that something akin to dependent variables (what this manual has been referring to with the broader label evaluation criteria) were used in analytical and judgment-based evaluations and surveys. No matter what the method used, an evaluation is conducted with certain evaluation criteria in mind. Within this sample, at least, it is worth noting that evaluations using experiment often use multiple evaluation criteria while those using other methods used a single criterion.

Table 7-1. Dependent Variables Used in SIMNET/CCTT Evaluations by Authors and Evaluation Methods

METHOD	AUTHOR (YEAR)	DEPENDENT VARIABLES
Experiment	Schwab & Gound (1988)	STX GO scores
	Brown, Pishel, & Southard (1988)	Platoon performance, command & control, and leadership
	TEXCOM (1990)	SME-rated performance on various collective tactical tasks
	Smith & Graham (1990)	Soldier performance on command and control (2) communications, (3) position location, (4) combat driving; rated similarity of training on SIMNET and with M1
	Hartley, Quillinan, & Kruse (1990a,b)	Conformance to direct fire and direct/indirect vulnerability models
	Shlechter, Bessemer, & Kolosh (1991)	Demonstrated leadership performance during simulated combat
	Bessemer (1991)	Amount and type of field training conducted, leadership performance
	Watson (1992)	Various tactical outcome measures; e.g., losses, exchange ratio, battle duration, kills
	Smith & Cross (1992)	Rated performance on individual and collective tasks and subtasks
	Shlechter, Bessemer, Nesselroade, & Anthony (1995)	Unit performance on training tables
	TEXCOM (1997)	Team performance on various tactical tasks
Analysis	Fusha (1989)	Trainability of selected MTP tasks
	Drucker & Campshure (1990)	Degree to which simulation supports training on task(s)
	Burnside (1990)	Degree to which simulation supports training on task(s)
	Thomas & Gainer (1990, May)	Degree to which simulation supports training on task(s)
	Noble & Johnson (1991a,b)	Adequacy of training on a family of tasks
	Lynn & Palmer (1991).	Degree to which simulation supports training on task(s)
	Scott, Djang, & Laferriere (1995)	Training effectiveness
	Finley (1997)	CCTT ability to simulate specified set of variations in communication quality
Judgment	Kraemer & Bessemer (1987)	Gunnery performance
	Brown & Mullis (1988a,b)	SIMNET realism and value for training
	Holstead (1989)	Relative capability of SIMNET to provide training on tactical aviation tasks
	Crane & Berger (1993)	Pilot interest in additional simulator training
	Hoffman (1997)	N/A. Resolution of various problems during introducton of new training program.
	Bessemer & Myers (1998)	Process performance indicators (determined on case-by-case basis)
Survey	Fletcher (1988)	Training value of SIMNET

Criteria and Small- Versus Large-Scale Evaluations

The evaluations shown in Table 7-1 vary in scale from small- to medium-sized. None is a full-blown evaluation of SIMNET/CCTT. Among the experiments, various single variables and combinations of variables were used. These include primarily measures of combat performance and collective task performance. Consider now the MDT2 evaluation, which could reasonably called an LSTS. The dependent variables used in this study are described in the sidebar.

A Concrete Example: Dependent Variables Used in the Multi-Service Distributed Training Testbed (MDT2) Project

The MDT2 project was conducted during 1994 and 1995 to test the feasibility of using virtual simulation to conduct multi-service training on the CAS (close air support) mission. Dependent variables used during the exercise included *Reaction*, *Collective Performance*, and *Results* measures.

Two types of collective performance measures were collected during exercises: TARGETs (targeted acceptable response to generated events or tasks), and TOMs (teamwork observation measure). Both measures were generated by SMEs who observed participant performance and recorded their observations using special data collection protocols. The TARGETs methodology is described as follows in Fowlkes, Lane, Salas, Franz, and Oser (1994):

> It is a form of structured observation in which (a) task events are introduced to provide opportunities for teams to demonstrate specific team-related behaviors; (b) acceptable team responses to each of the events are determined a priori by utilizing team task analyses, subject-matter experts, and so forth; and (c) the appropriate responses to events are scored as either present or absent (p. 47).

TOMs data reflect the adequacy of interactions among team members (i.e., Service representatives) for each of three mission phases (planning, contact point, attack) and four dimensions (communication, coordination, adaptability, situational awareness). TOM was intended to identify strengths and weaknesses in teamwork (Dwyer, Fowlkes, Oser, and Salas, 1996).

Results measures were obtained using the UPAS (unit performance assessment system). UPAS was developed by the Army to calculate and display performance measures and summary statistics associated with SIMNET exercises. UPAS gathers data from five sources (network, terrain, unit plans, radio communications, direct observations) and generates information on vehicle appearance, status, and status change and fire, indirect fire, and impact. The UPAS data were recorded during each exercise, permitting later playback to develop these *results* measures:

- Number, timing, and frequency of bombs released by F-16s
- Number of vehicles hit, damaged, or destroyed
- Percentage of bombs resulting in a vehicle impact or near impact
- Number of bombs causing damage or destruction
- Timing and volume of artillery direct fires and CAS fires
- Timing and location of direct and supporting fire impacts

Reaction measures were gathered in a written survey conducted at the conclusion of the exercise (Mirabella, Sticha, and Morrison, 1997). All exercise participants and O/Cs completed a written survey to give their opinions and comments on how well MDT2 had worked and what value it added.

The MDT2 evaluation used a wider range of dependent variables than the smaller-scale evaluations conducted for SIMNET/CCTT. This is not surprising. A small-scale evaluation may use one or two narrow dependent variables to answer whatever limited question it addresses. On the other hand, a large-scale evaluation must use a range of variables to answer the much broader question it addresses.

Criteria and Evaluation Perspective: Training Versus System Developer Versus M&S

As noted in Chapter 2, LSTS are sufficiently complex and costly that DoD acquisition regulations lay out an orderly succession of developmental phases:

- Phase 0: Concept exploration
- Phase I: Program definition and risk reduction
- Phase II: Engineering and manufacturing development low rate initial production
- Phase III: Production, fielding/deployment, and operational support

Milestone decision points, established early in the program, determine whether or not the program is progressing satisfactorily and may proceed to the next phase. It is within this overall process that training evaluators work. System developers and evaluators have potentially competing interests and conflict can arise based on how progress is measured.

How, exactly, is progress supposed to be measured? Guidance on this subject is found in *DoD Regulation 5000.2-R: Mandatory Procedures for Major Defense Acquisition Programs (MDAPs) and Major Automated Information System (MAIS) Acquisition Programs* (Department of Defense, 1996b), which states the following about what to measure during program development tests:

> At Milestone I, performance parameters shall be defined in broad terms. Measures of effectiveness or measures of performance[76] shall be used in describing needed capabilities early in a program. More specific program parameters shall be added as necessary.... The total number of performance parameters shall be the minimum number needed.... This minimum number shall include the key performance parameters described in the ORD [operational requirements document].... These performance parameters may not completely define operational effectiveness or suitability. Therefore, the MDA[77] may add additional performance parameters.... (part 3, page 2).

This guidance is general and gives much discretion to the evaluator in terms of what to evaluate and what to use as dependent measures. It appears that the evaluator can settle for the minimum set of performance parameters or, if so inclined, be more ambitious. The usual concern is that program managers will focus on hardware and software and not on training effectiveness. Alert

[76] Measure of Effectiveness (MOE) and Measure of Performance (MOP) refer to two classes of dependent measures commonly used in the M&S community. They are defined as follows in *DoD Directive 5000.59-M (Glossary of Modeling and Simulation Terms)* (Department of Defense, 1998). MOE: A qualitative or quantitative measure of the performance of a model or simulation or a characteristic that indicates the degree to which it performs the task or meets an operational objective or requirement under specified conditions. MOP: Measure of how the system/individual performs its functions in a given environment (e.g., number of targets detected, reaction time, number of targets nominated, susceptibility of deception, task completion time).

[77] MDA (Milestone Decision Authority) is the individual with authority to approve entry of an acquisition program into the next phase (Department of Defense, 1996b).

readers should at this point be getting a little alarmed. The regulation is vague and does not mention training effectiveness. To assure that this interest is represented, someone must advance it and act as its proponent.

One other complication in evaluating LSTS is that they are classified as major M&S developments or upgrades and are required to go through a process called VV&A (Verification, Validation, and Accreditation) as described *in DoD Instruction 5000.61 (DoD Modeling and Simulation [M&S] Verification, Validation, and Accreditation [VV&A])* (Department of Defense, 1995) and the Defense Modeling and Simulation Office's *Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide* (Department of Defense, 1996c). DoD Instruction 5000.61 defines these activities as follows:

- Verification: The process of determining that a model implementation accurately represents the developer's conceptual description and specifications.
- Validation: The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended users of the model.
- Accreditation: The official certification that a model, simulation, or federation of models and simulations is acceptable for use for a specific purpose.

The exact timing of VV&A in relation to training effectiveness evaluation is not specified in acquisition policy; this factor has not been recognized as a problem. However, VV&A addresses questions related to training effectiveness evaluation. Having the proponents of training effectiveness evaluation and VV&A ignore what each other is doing is analogous to building an aircraft by hiring separate contractors to work on the major systems (e.g., airframe, flight control, electronics, etc.) without coordinating their activities. Likewise, bring into the equation the interests of a third interested party, the program manager, and there is ample opportunity for conflict.

Conflicts among vested interests may be fought out as political games with winners and losers. A capable and enlightened program manager may be able to circumvent this by coordinating the activities of the various interested parties. There are no formal guidelines or mechanisms for doing this at present. Each new evaluation presents a new opportunity to invent one.

Evaluation Criteria for an LSTS Evaluation

Industrial Training Program Evaluation

Kirkpatrick's (1976) framework for evaluating training programs has gained wide acceptance in the industrial training community. An LSTS is not an industrial training program, although the two have in common that both train adult learners in job skills for the purpose of improving job performance in the workplace.[78] Kirkpatrick recommends that training programs be evaluated at four levels: reaction, learning, behavior, and results. Data gathered at each of these levels can answer different questions about the effect of the training program on its students:

- Reaction: How well did students like the program?
- Learning: What did students learn while participating in the program?
- Behavior: What changes in job behavior resulted from the program?
- Results: What were the tangible results after the program in terms of reduced cost, improved quality, improved quantity, and so forth?

Note that (1) reaction and learning data are gathered during training and (2) behavior and results are gathered after training. Typical ways to gather data for each of these levels for reaction would be post-course survey; for learning, in-course tests; for behavior, post-course supervisorial performance evaluations; for results, student post-course productivity and work quality. Now, consider these categories from the perspective of an LSTS.

Reaction

The simulation is a complex, collective learning environment in which O/Cs evaluate the performance of teams and other collectives. Reactions of trainees to their learning experience are important. Reactions of O/Cs are also valuable. Further, it is sometimes useful to gather reaction data on other questions; for example, preferences for certain design features, suggestions for changes, and so forth. Conclusion: Expand the scope of this category to include the additional reaction variables indicated.

Learning

The Learning of importance in the simulator is collective learning. Learning is reflected in a change in collective performance with time. In the classroom, individual learning might be measured by changes in test scores from before to after training. In the simulator, collective learning is reflected in improved performance of the collective.[79] Conclusion: Rename this category Collective Performance.[80]

[78] Mirabella (1998, July 31) notes that Kirkpatrick's model was designed for "industrial, classroom, lecture/content oriented, individual training." Alternatively, training with LSTS occurs in a simulated workplace (e.g., combat vehicle), training is delivered through combat exercises, and the focus is on collective training.

[79] In general, collective performance is measured by O/Cs who observe the process of interactions among members and record key incidents using data collection protocols. As this process improves, performance is said to have improved, and learning to have occurred. While analogous, the concept of individual learning in the classroom does not map directly to collective learning in the simulator.

[80] Large-scale training simulations are intended to provide training on collective tasks. At the Joint level, these tasks are listed in the Universal Joint Task List (UJTL). Service-specific task lists define the relevant collective tasks at the Service level. These task lists essentially lay out what tasks the Services and Joint forces are expected to be able to perform. They are the logical tasks to use when building scenarios to evaluate LSTS. In other words, collective performance should be evaluated based on these lists.

Behavior

The post-training behavior of importance is collective performance. This category is equivalent to the Collective Performance category, immediately above, except that it measures performance after training has occurred. Conclusion: Rename this category Post-Training Collective Performance.

Results

The post-training results of interest for the simulator are the military readiness and combat results the simulator was intended to support; for example, military readiness, and simulated and actual combat outcomes as reflected in such variables as exchange ratio, percent losses by force, shots/kill, and so forth. Conclusion: Expand the scope of this category to include the military variables indicated.

This analysis leaves four slightly redefined and renamed categories: Reaction, Collective Performance, Post-Training Collective Performance, and Results.

What Military Training Experts Have Recommended

A number of training and evaluation experts who have evaluated LSTS have recommended that certain dependent measures be used. These recommendations should be considered before developing a list based entirely on Kirkpatrick's framework. The Kirkpatrick framework was not developed to evaluate LSTS, although it appears to be suitable for this purpose with some modifications. To validate this impression, consider what the experts have recommended and see how well they fit within the Kirkpatrick-based framework described in the previous section.

Alluisi (1991) makes the case that post-training readiness is a relevant dependent measure.[81] Bell, Dwyer, Love, Meliza, Mirabella, and Moses (1997a) recommend that evaluators measure both system processes (such as interactions among team members) and combat outcomes in the simulator (such as casualties by weapon system, loss exchange ratios, and amount of terrain controlled). They also recommend that players express their judgments about how well training objectives were satisfied. Garlinger and Fallesen (1988) recommend that evaluators focus on (1) user acceptance, (2) processes, and (3) achievement of system goals. Hiller (1997) recommends that evaluators focus on (1) task and mission performance outcomes, (2) task and task step performance processes, and (3) user and SME comments regarding simulation features. Hiller (1994, 7 February) also recommends the use of archival data to estimate post-training effects on job performance. Hiller's recommendation is consistent with those made by Boldovici and Bessemer (1994) and Leibrecht (1996).

[81] Alluisi made this case regarding the evaluation of SIMNET: "For SIMNET to be viewed by the Army as successful—that is, as an effective training system that is worth the expenditure of funds for additional procurement—the Army will have to be convinced that it will make a difference in readiness" (p. 360).

Most of the dependent measures recommended by these authors fit into the four categories derived from Kirpatrick. However, three of these authors refer to an in-simulator equivalent of Kirkpatrick's Results category; these are Bell et al's "combat outcomes in the simulator," Garlinger and Fallesen's "achievement of system goals," and Hiller's "task and mission performance outcomes." These are peacetime, in-simulator measures of performance in simulated combat. Call them During-Training Results.

Terminology differs, but the consensus is that evaluators should use a combination of reaction, collective performance, and results measures. Table 7-2 is an attempt to sort out the dependent measure "votes" by author. These modify and extend Kirkpatrick's list to make it more suited for LSTS evaluation.

Table 7-2. Consolidated List of Recommended Dependent Measures by Author

WHEN	DEPENDENT MEASURE	AUTHOR			
		Alluisi	Bell et al.	Garlinger & Fallesen	Hiller
During training	1. Reaction		√	√	√
	2. Collective Performance		√	√	√
	3. Results		√	√	√
Post-training	4. Collective Performance				√
	5. Results	√			√

On the basis of the dependent measures forwarded in Table 7-2, Table 7-3 presents a consolidated list of dependent measures with descriptions adapted from Kirpatrick.

Table 7-3. Consolidated List of Recommended Dependent Measures with Descriptions

WHEN	DEPENDENT MEASURE	DESCRIPTION
During training	1. Reaction	What were user and O/C reactions to simulator?
	2. Collective Performance	How well did teams and other collective echelons *perform* in the simulator?
	3. Results	What were the *tangible results* during training? (exchange ratio, percent losses by force, shots/kill, etc..)
Post-training	4. Collective Performance	How did team and other collective echelons perform after training?
	5. Results	What were the *tangible results* after training? (readiness, field exercise performance, combat outcomes)?

This scheme uses five different classes of variables. Note that:

- The first three (Reaction, Collective Performance, Results) are obtained in the training system. The last two (Collective Performance, Results) are obtained post-training.
- Variables 2 and 4, and 3 and 5 are analogous pairs, with the first number reflecting performance during training and the second post-training.

Think of Table 7-3 as a shopping list of measures to consider during evaluation. Each of these is a class of measures rather than a single measure. Each could be represented in several different ways. For example, in any single form, Collective Performance reflects the performance of a collective at a particular level in an organizational hierarchy. The most basic level would be the team. Collective Performance could also be measured at higher echelons; for example, squadron, wing, battalion, brigade, division, corps, theater army, multi-service force, joint force, and so forth. In an actual evaluation, more than one type of Collective Performance might be measured. This also holds true for the other four classes of measures.

The measures are not all of equal significance. Reaction data are useful, but less important than Collective Performance, which itself is less important than Results in the simulator. None of these is as important as performance in the real world, which means that the post-training measures are the most important of all. One form of post-training Results is to win or lose a battle. Unfortunately, it usually gets more costly and difficult to obtain measures as importance increases.

Gathering information across the range of measures gives a better understanding of how well training is working. If the only measure used were Results, and Results were poor, it would be difficult to diagnose the underlying cause. It is important to gather data on lower-level and intermediate measures.

What measures should be used to evaluate an LSTS? This varies with the situation. The sidebar earlier in this chapter gave concrete examples of the types of Reaction, Collective Performance, and Results measures used in the MDT2 evaluation. Table 7-1 lists various measures used in the SIMNET/CCTT evaluations. The classes of measures derived in this chapter and the examples given should help the reader define appropriate evaluation criteria for new evaluations.

8 EVALUATION FRAMEWORK

This chapter further develops the training effectiveness evaluation framework introduced in Chapter 2. Chapters 2-7 provide background information necessary to understand the present chapter. Readers should review those chapters before attempting to apply the ideas in the present chapter in an evaluation.

An evaluation framework is a set of evaluation principles and a description of evaluation events, their purpose, timing, and relevant dependent variables. Regard the framework as suggestive rather than prescriptive. It is intended to help the evaluator select the most suitable evaluation methods based on the circumstances, provide procedural descriptions of the methods, and identify case studies; that is, examples of completed studies linked to each method that can be used as models to emulate.

The chapter begins by describing a set of evaluation objectives. It then presents a set of evaluation principles that comprise an evaluation philosophy. Then, in the section titled Evaluation Framework, it ties together evaluation objectives, events, and methods on a timeline such that the resulting process is true to the declared evaluation principles.

Reference List 8-1 (Evaluation Framework) at the end of this chapter contains complete citations for publications cited in this chapter.

Evaluation Objectives

The logical beginning of an evaluation is to define its objectives. Evaluation objectives were discussed briefly in Chapter 2. This section will attempt to bring the discussion into closer focus by defining a limited set of objectives that apply during LSTS evaluation. The evaluator must decide which of these applies in a particular situation. Note that an evaluation may be conducted with more than one objective in mind. To simplify discussion in what follows, define the time window prior to and including MDAP Phase 0 as pre-development, Phases I and II as developmental, and Phase III as post-development. Table 8-1 summarizes some of the most common objectives for conducting evaluations.

Table 8-1. Common Objectives for Conducting Training Effectiveness Evaluations

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
1	Estimate need for new training system	N/A	
2	Predict training effectiveness	Measure training effectiveness	Determine training effectiveness
3	Predict transfer of training	Measure transfer of training	Determine transfer of training[1]
4	Predict user acceptance	Measure user acceptance	Determine user acceptance
5	Support training design[2]		Determine training status
6	Support system design[3]		Evaluate system design

[1] Includes (A) determine effects of training on performance, (B) readiness, (C) use of resources.

[2] Includes (A) design training, (B) identify training problems, (C) resolve training problems.

[3] Includes (A) assure adequate learning environment and (B) data collection, analysis, retrieval, and display for AAR.

Before system development begins, a decision is made to start development. This decision may be based on a study to estimate the need for a new training system. This objective is represented by the row labeled Code 1 in Table 8-1.

By far the most common evaluation objective is to predict, measure, or determine training effectiveness (Code 2). Most milestone evaluations are conducted to satisfy this objective.[82] This objective actually consists of three sub-objectives:

- 2A. Predict training effectiveness (pre-development): estimate effectiveness before the training system is operational.
- 2B. Measure training effectiveness (developmental): estimate effectiveness during system development.
- 2C. Determine training effectiveness (post-development): integrate data post-development to reach definitive conclusions about training effectiveness.

Evaluations may be used to predict, measure, or determine transfer of training (Code 3). Of particular interest is transfer of training from the simulator to settings that reflect, in varying degrees, performance in wartime; for example, field training or live simulation training. Also of interest here is the effect of training on unit readiness. This objective consists of three sub-objectives, analogous in timing to those for Objective 2.

Evaluations may be conducted to predict, measure, and determine user acceptance (Code 4). This objective consists of three sub-objectives, analogous in timing to those for Objectives 2 and 3.

Evaluations may be conducted to support training design (Code 5); for example, to select among alternative training strategies. Studies may be conducted to identify and correct training problems. Post-development, evaluations may be conducted to determine training status; for example, how well individuals in a particular MOS (Military Occupational Specialty) are able to perform their jobs.

[82] One of the most obvious reasons to conduct evaluations is to *satisfy milestone requirements.* This is a reason but not an evaluation objective; the usual objective in milestone evaluations is to measure training effectiveness (sub-objective 2B). As noted in Chapter 7, the guidance on how to conduct milestone evaluations is general and gives considerable discretion to evaluators in terms of what to evaluate and what to use as dependent measures. Milestone evaluations may be a big deal—large in scale, costly, and with many participants. On the other hand, they do not have to be. It all depends upon what are set as the evaluation objectives based upon an interpretation of required capabilities as defined in acquisition documents. Many of these evaluations are limited in scope, particularly during the early phases of training system development.

Evaluations may be conducted to support system design (Code 6); for example, to assure that the design provides an effective learning environment. After development is complete, the design may be further evaluated.

Evaluation Principles

The evaluation framework is based on a set of evaluation principles, which represent the philosophy adopted toward evaluation. Most of these principles should already be obvious to the reader based on what has come earlier in this manual and common sense. Keep in mind that this is not a procedure and is not meant to be followed in step-by-step fashion.

Determine Evaluation Stakeholders

Stakeholders are those with a vested interest in evaluation. They vary with circumstances, but may include program managers, developers, training evaluators, military decision-makers, and others. Stakeholders must cooperate to make an evaluation succeed. The first step in any evaluation is to determine who these stakeholders are. The second step is to determine what information, obtained during evaluation, will satisfy each stakeholder.

Define Objectives

An evaluation requires clearly-defined objectives at the outset. An evaluation may be conducted with more than one objective in mind; for example, to satisfy a milestone requirement while simultaneously demonstrating training effectiveness. Further, there must be consensus among stakeholders on evaluation objectives.

Treat Evaluation As a Process, not an Isolated Event

As previously discussed, evaluations are often thought of as one-shot events that answer a question at a particular point in time. This makes little sense when evaluating complex and expensive LSTS that undergo years of development before becoming operational. LSTS evaluation may occur as a series of evaluation events, culminating periodically in larger milestone events, and eventually in a live or die Phase III evaluation. Given that evaluation cannot be done in a single stroke, the question becomes one of developing a logical progression of events that will support the development and fielding of an LSTS with the greatest possible training effectiveness. One of the best examples of how this process may unfold is in the SIMNET case study, presented in Chapter 4, in which more than two dozen evaluation events occurred.

Attempt To Influence Design and Development

Training experts should play an important role in the design and development of training systems. Historically, this has not always been the case. Training evaluators should be brought into the system design process to influence system design from a learning perspective; that is, to assure that the design provides an adequate learning environment. There appears to have been a disconnect between system developers and the training community. In the JSIMS development, the Navy and OSD have supported a formal effort by the JSIMS Learning Methodology Working Group (LMWG) to influence system design from a learning perspective.[83]

Evaluate Multi-Dimensionally

In the first of the *Back to the Future* movies (Universal City Studios, Inc., 1985), the character played by Christopher Lloyd ("Doc") attempts to overcome the confusion of the Michael J. Fox character ("Marty") about the complications of time travel by advising him to try four-dimensional thinking. Similar advice might be offered to training evaluators contemplating LSTS evaluation. The four dimensions the evaluator needs to link together in the mind are (1) evaluation objectives, (2) time, (3) evaluation criteria (dependent variables), and (4) evaluation methods. Table 8-1 already linked together the first two of these four dimensions by illustrating how objectives may change as a function of time or stage of system development.

The third dimension, evaluation criteria, may be added to this pair by considering that different sets of dependent variables may be used depending upon the evaluation objective. To illustrate, contrast Objectives 2 and 4 at the developmental stage in Table 8-1. To measure compliance with Objective 2 (Measure training effectiveness) the evaluator would be well advised to use the full set of dependent variables developed in Chapter 7 (Reaction, Collective Performance, and Results) during training. To measure compliance with Objective 4 (Measure user acceptance), it is enough to gather Reaction data alone.

The fourth dimension, evaluation methods, may be added to this triad by considering the logical types of evaluation methods needed to collect the dependent variables. For the effectiveness evaluation, this represents a combination of methods; for example, experiment, judgment, and possibly survey. For evaluating user acceptance, judgment and/or survey methods would make sense.

[83] The LMWG was formed because JSIMS development priorities are weighted heavily toward technical engineering needs rather than being balanced with the training and learning perspective advocated by the behavioral sciences and user communities (Learning Methodology Working Group, 1999).

Obtain the Best Data Possible

The worth of an evaluation depends upon the quality of its data in terms or relevance, validity, and reliability. Beware the fallacy that one evaluation method is inherently superior to another. The quality of data obtainable with a particular method may outweigh other considerations. Beware the common pitfalls noted by Boldovici (1987), Boldovici and Bessemer (1994), and Kraemer and Rowatt (1994) and cited in Chapter 5; for example, not enough subjects, differences between compared groups, different treatment of groups, insufficient amount of practice to affect proficiency, ceiling and floor effects, unreliable test scores, untimely administration of transfer tests, use of inappropriate analyses, and misinterpretation of null results.

Develop Learning Curves

If a training event can be repeated several times during an evaluation, it may be possible to develop learning curves. The curves show not only that learning occurred or did not occur, but the rate of learning across time. Learning curves are more informative than point measures in determining the course of learning. They are most readily developed when conducting experiments. They are particularly useful when the situation precludes the use of a control group because they permit inferences about learning. Good examples of such curves are presented in Figures 4-1 and 4-2 for the MDT2 evaluation. Good examples of learning curves for judgment data are presented in Wetzel, Simpson, and Seymour (1995).

Measure Transfer of Training

Transfer experiments measure the effects of learning in one situation to performance in another. Obviously, the greater the amount of transfer that can be demonstrated to the combat environment, the more convincing the evidence. Chapter 3 described three different types of transfer experiments (Validation, Comparison, and Relationship). While all of these can be useful, the transfer of primary concern during LSTS evaluation is validation; that is, demonstrate transfer from training system to the job. Experiments are not the only way to estimate transfer. It may be possible to use an analytical method (such as simulated transfer—see Chapter 3). System users and SMEs may also be asked to estimate transfer in judgment-based evaluations and surveys.

Evaluation Framework

This section lays out an evaluation framework that ties together evaluation objectives, criteria, and methods consistent with the evaluation principles declared earlier. The section is organized based on the six classes of evaluation objectives in Table 8-1. (The rationale underlying each of these objectives was described under Evaluation Objectives, earlier in this chapter, and will not be reprised here.) Evaluation criteria and methods to satisfy each objective are discussed and representative examples[84] of completed evaluations are summarized.

1. Estimate Need for New Training System

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
1	Estimate need for new training system	N/A	

Before system development begins, a decision is made to start a development. This decision may be based on a study to estimate the need for a new training system. Such estimates are typically based on analysis or judgment/survey. Evaluation criteria would be estimates of need for the new system; for example, ratings on a scale from 1-10. An analytical study might use an estimate of cost-effectiveness as the evaluation criterion.

TCEF includes two studies that fall into this category.

- Bretl, Rivera, and Coffey (1996). Study was conducted to determine need for, characteristics of, and cost of hypothetical ENCATT (Engineer Combined Arms Tactical Trainer) Soldiers from combat engineer units and the Engineer School completed written survey instruments with questions relating to need for ENCATT, tradeoffs, and frequency and importance of training on collective tasks. A cost estimate was developed. Includes study plan and data collection survey instruments. Evaluation criteria: estimated cost and training effectiveness. Evaluation method: analysis (evaluate).
- McDade (1986). Prospective evaluation of a hypothetical simulator to train BFV (Bradley Fighting Vehicle) drivers. Study objective: determine need for driver trainer and if it would be a cost-effective way to train Bradley drivers. Driving tasks were identified. Driver training effectiveness was assessed by observing training in schools and units and by administering questionnaires and interviews to commanders, supervisors, instructors, and Bradley crews. Driver training costs were estimated with and without simulators. Conclusion: driver trainer would not be cost-effective. Evaluation criteria: estimated cost, and frequency and importance of training tasks. Evaluation method: survey and judgment (users).

[84] In selecting examples for Tables 8-2 through 8-10, care was taken to select studies that reflect the distribution of evaluation criteria and methods within the particular subsample of studies for the objective. However, because of the small sample size for some of these objectives (e.g., 1, 4, and 6) representativeness is questionable.

2. Predict, Measure, or Determine Training Effectiveness

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
2	Predict training effectiveness	Measure training effectiveness	Determine training effectiveness

Nearly 80 percent of the evaluations in TCEF (including all of its milestone evaluations) were conducted to satisfy one of the three sub-objectives of Objective 2. This is the most common reason to evaluate training.

2A. Predict Training Effectiveness.

Table 8-2 provides a brief descriptive summary and gives the evaluation criteria and method for a sample of evaluations for Sub-objective 2A. These evaluations were conducted pre-development. All but one of these evaluations were analytical; Kelly's was judgment-based. The first three listed were milestone evaluations. Most of the evaluation criteria were task-related estimates of how well each training system was able to train on a particular set of tasks. No actual performance data were collected in any of the evaluations.

2B. Measure Training Effectiveness

Table 8-3 provides information for a sample of evaluations for Sub-objective 2B. These evaluations were conducted during training system development. Evaluations conducted to meet this sub-objective are the most numerous of all evaluations in TCEF. Virtually all of the evaluations presented as case studies in Chapter 4 fall into this category. The distribution of evaluation criteria and methods for the sample of evaluations in Table 8-3 approximates that for this objective in TCEF as a whole. The majority of these evaluations used experiment. Judgment and analysis were also used, but less frequently.

Table 8-2. Descriptive Summary of Representative Evaluations for Evaluation Objective 2A: Predict Training Effectiveness

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Carroll (1995)	Training effectiveness and cost	analysis (compare)	Milestone evaluation. Objective was to determine the most cost-effective training strategy for Heavy Assault Bridge, a longer version of Breacher. This study extrapolated from the earlier Breacher CTEA. Breacher CTEA was analyzed to identify bridging specific tasks and training alternatives were generated; these were reviewed by SMEs. Training methods and resources were estimated. Alternative training strategies were developed. Costs were estimated for the alternative strategies. Sensitivity analysis was conducted. Training strategy was determined by comparing relative costs and estimated effectiveness of alternatives. Methods described in detail.
Noble & Johnson (1991a,b)	Adequacy of training on a family of tasks	analysis (compare)	Milestone evaluation. Analytical study to determine possible OPTEMPO reductions with adoption of CCTT. CCTT training effectiveness was estimated based on previous analyses of SIMNET (surrogate system). CCTT TDR was examined to determine task areas to be trained; these were compared with task areas covered by SIMNET. Three different training device alternatives were compared (improved SIMNET-T, degraded CCTT, embedded training). Costs were estimated.
Leatherwood, Schisser, & Russell (1986)	Training effectiveness and cost	analysis (evaluate)	Milestone evaluation. A CTEA for a training program that had not yet been tried at the time of the study. A task list was developed based on documentation, site visits, and related courses. The list was reviewed by a panel and revised. POIs were reviewed and critiqued in relation to task coverage. The envisioned courses were found to be inadequate. What, exactly, happened in this study is somewhat ambiguous.
Finley (1997)	CCTT ability to simulate specified set of variations in communication quality	analysis (evaluate)	Prospective evaluation of the capability of the CCTT to provide a suitable environment for training involving degraded communications. Analyses were performed to first identify training needs in armor and mechanized infantry units using single channel ground/air radio systems. Capabilities of initial CCTT to simulate realistic variations in communications quality were then estimated.
Berg, Adedeji, & Trenholm (1993)	Marksmanship performance	analysis (evaluate)	An analytical study of the potential use of simulators vs. live-fire for USMC marksmanship training. The study examines simulators currently used in the USMC and Army and their estimated potential for expanded use in the USMC to effect cost savings. The biggest cost driver in marksmanship training is the cost of training ammunition. Additional costs are involved in operating and maintaining ranges. Simulators have the potential to significantly reduce these costs in the USMC. Limited effectiveness data are provided. Detailed cost analyses are provided.
Lynn & Palmer (1991)	Degree to which simulation supports training on task(s)	analysis (evaluate) judgment (analysts)	Analysts reviewed various CCTT conceptual documents (concept evaluation program; training device needs statement, training device requirement, system specification) and reports (reliability; force development test and experimentation final report) and estimated operational effectiveness of CCTT. CCTT strengths and weaknesses were extrapolated from those of SIMNET.
Drucker & Campshure (1990)	Degree to which simulation supports training on task(s)	analysis (evaluate)	An analysis to estimate how well SIMNET can be used to train tactical activities conducted during tank platoon operations. The activities performed by armor personnel during combat were identified from field manuals and other documents. The research staff then attempted to perform these activities on SIMNET and recorded estimated fidelity with a checklist.
Burnside (1990)	Degree to which simulation supports training on task(s)	analysis (evaluate)	SMEs rated degree to which selected ARTEP tasks could be performed in SIMNET. Ratings were consolidated with decision rules, reviewed, and coordinated.
Kelly (1995)	Functional training capabilities of simulator	judgment (SMEs)	SMEs separately rated training capabilities of traditional method (Range 400) and Leathernet (pre-build system).

Table 8-3. Descriptive Summary of Representative Evaluations for Evaluation Objective 2B: Measure Training Effectiveness

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Thomas & Gainer (1990, May)	Degree to which simulation supports training on task(s)	analysis (evaluate) judgment (SMEs) judgment (users)	Case study to evaluate how well AIRNET could be used to train ARTEP tasks. Tasks were selected. Pilots used AIRNET to conduct simulated missions. SMEs rated their performance and AIRNET performance for each task. Subjects completed questionnaires about technical performance of system.
Orlansky, Taylor, Levine, & Honig (1997)	Reactions, adequacy of team interactions, bombing performance	experiment (quasi-) survey	Cost and training effectiveness evaluation of the MDT2, a prototype virtual simulation for training the close air support mission and involving multi-service aid and ground forces. Process and outcome measures were obtained on a daily basis during 5-day exercise. Participant judgment data were obtained at end of exercise. MDT2 was effective in training, and cost approximately one-tenth of equivalent training using live forces. The methods are described in sufficient detail to be useful as models by evaluators.
Shlechter, Bessemer, Nesselroade, & Anthony (1995)	Unit performance on gunnery training tables	experiment (quasi-) judgment (SMEs) judgment (users)	Unit scores were obtained and compared across six successive gunnery training tables. Instructors and participants provided ratings of the training experience.
Smith & Cross (1992)	Rated performance on individual and collective tasks and subtasks	experiment (test)	Aircrews performed a variety of individual and collective tasks on simulator and their performance was rated by SMEs; aircrews also completed questionnaire items
Lickteig & Collins (1995)	Numerous, based on blueprint of battlefield; e.g., loss/kill ratio, % kills, no. hits, hit range, kill range, hits/round ratio, kills/hit ratio, kills/round ratio, no. rounds fired	experiment (true)	2 x 3 factorial experiment (CVCC and baseline conditions by battalion, company, and platoon echelons) between-subjects design. Objective was to determine operational effectiveness of CVCC connectivity among exercise participants. Baseline groups underwent similar training but were not equipped with CVCC. Report describes method in sufficient detail to use as model in comparable TEAs involving large-scale simulations and/or field training.
Shlechter, Kraemer, Bessemer, Burnside, & Anthony (1996)	SME attitudes toward various aspects of VTP	judgment (SME) survey	Survey of SMEs (observer controllers) regarding their attitudes toward the virtual training program in terms of these aspects of VTP: train the trainer, unit preparation, training structure, training proficiency, unit follow up and take home packages. Participants were interviewed about selected aspects of VTP.

Table 8-4. Evaluation Criteria by Study for Objective 2B: Measure Training Effectiveness

WHEN	EVALUATION CRITERIA	STUDY					
		Thomas & Gainer	Orlansky et al.	Shlechter et al. (1995)	Smith & Cross	Lickteig & Collins	Shlechter et al. (1996)
During training	1. Reaction	√	√				√
	2. Collective Performance		√		√		
	3. Results		√	√		√	
Post-training	4. Collective Performance						
	5. Results						

Table 8-4 breaks down the evaluation criteria based on the classes of variables developed in Chapter 7. All of these studies used at least one of the variables presented in Table 7-3. The Orlansky et al. study (i.e., the MDT2 evaluation) used all three of the During Training variables. This study arguably represents the best model to emulate in LSTS evaluations published to date.

2C. Determine Training Effectiveness

Table 8-5 provides information for a sample of evaluations for Sub-objective 2C. These evaluations were conducted post-development. This sample has limitations for generalization to LSTS. Only the Orlansky et al. and Worley et al. evaluations deal with LSTS. These two studies are retrospective reviews of the literature relating to LSTS training effectiveness. The Bailey and Hodak evaluation deals with weapons simulators. The remaining evaluations are for training programs. There are not many retrospective evaluations of LSTS in TCEF. Despite this limitation, the evaluations are informative. First, all are based mainly on existing data; for example, an audit trail for a training course, a set of published studies, or school records. The underlying data were reviewed and compiled to draw generalizations about training effectiveness. Second, evaluation methods were primarily analytical. An exception is the Derrick and Davis evaluation, which used an ex post facto experiment based on comparative data from two different training programs.[85]

Would it make sense to use an experiment to determine training effectiveness? It would if an experiment could be conducted that definitively answered the question. Usually the question is approached more conservatively by conducting reviews such as Orlansky et al and Worley et al. Experiments might conceivably provide such definitive data, though it is more common practice to evaluate one step at a time to build up a convincing body of evidence.

[85] This study has been widely cited and often praised. Its methods section is sufficiently detailed that it could be followed in conducting a comparable study. The study draws together convincing evidence to make the case that contractor training is more economical than training conducted by military personnel. Caveat: the ability to conduct a study such as this depends on the existence of detailed, long-term historical records.

Table 8-5. Descriptive Summary of Representative Evaluations for Evaluation Objective 2C: Determine Training Effectiveness

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Ambruster (1987)	Student critiques, flight & academic grades, instructor interview comments	analysis (evaluate)	A review of two related ongoing training programs. SMEs reviewed the class "audit trail", student critiques, flight evaluation grades, comment slips, academic results, training materials, PIs, and interviewed instructors.
Bailey & Hodak (1994)	Marksmanship accuracy	analysis (evaluate)	A review of several studies evaluating effectiveness of weapon simulators: Multipurpose Arcade Combat Simulator, Weaponeer, Squad Engagement Training System, Indoor Simulated Marksmanship Trainer, Precision Gunnery Training System. Also sketches some methods for evaluating live fire offset. Though some results are positive in terms of reduced costs and transfer from simulator to live fire, study concludes that simulation offset to live fire could not be determined at time of study because of lack of empirical data and need to rely on perceptions.
Orlansky, Dahlman, Hammon, Metzko, Taylor, & Youngblut (1994)	Training effectiveness and cost	analysis (evaluate)	A wide-ranging review of the cost and effectiveness of military models and simulations as they relate to training. Estimates investments in M&S, summarizes cost-effectiveness findings, describes M&S usage by service, describes distributed interactive simulations in use, and sketches key technologies relevant to simulation and training (e.g., networks, semi-automated forces, range instrumentation, dismounted combatants, virtual environments, etc.)
Worley, Simpson, Moses, Aylward, Bailey, & Fish (1996)	Training effectiveness and cost	analysis (evaluate)	Review of the literature on the cost and training effectiveness of simulation at several training echelons (individual, collective, command and staff). Demonstrates cost-effectiveness of simulation. Provides comparable review for acquisition and analysis applications of modeling and simulation.
Derrick & Davis (1993)	Training effectiveness and cost	experiment (ex post facto)	Comparative study of large training system comprising 43 courses taught to pilots, navigators, flight engineers, loadmasters, and maintenance technicians. Study compared the costs and effectiveness of traditional aircrew training system (conducted by USAF personnel) and contractor-delivered (flying training delivered by USAF). Training effectiveness for two programs was assessed by examining training folders for both training periods. Cost data were obtained by counting resources for both systems; e.g., number of graduates, instructors, airplanes, flying hours, training days, overhead staff, types and number of training devices, etc. Method described in detail.
Evans & Braby (1983)	Student, instructor, and supervisor ratings of instructional quality	judgment (users)	Survey of 37 Navy and Marine Corps courses involving individualized instruction. Data were collected from site visits, reviews of materials, and questionnaires administered to students, instructors, and supervisors. Conclusions: perceptions of individualized instruction were positive but quality of conventional instruction was rated higher. Cost of conventional instruction was higher.

3. Predict, Measure, or Determine Transfer of Training

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
3	Predict transfer of training	Measure transfer of training	Determine transfer of training

The three training transfer Sub-objectives (3A, B, C) are analogous to those for training effectiveness (2A, B, C); namely, to predict, measure, and determine transfer of training. All of the studies cited in this section are for the aviation and gunnery training content areas. None of these studies deals with LSTS. While evaluators often cite the benefits of transfer studies to establish training effectiveness, such studies are rarely conducted.

3A. Predict Transfer of Training

No studies in TCEF were conducted to satisfy this objective. It would be nice to be able to predict transfer of training, but no method to do this is widely accepted and used. Analytical methods can be used to make these predictions. Refer to discussions of DEFT, FORTE, Simulated Transfer, and Comparison-Based Prediction in Chapters 3 and 6.

3B. Measure Transfer of Training

Table 8-6 provides information for a sample of evaluations for Sub-objective 3B. Most of these evaluations were conducted with operational simulators. All of the evaluations used experiment. The evaluations in the top three rows of Table 8-6 are for gunnery and in the bottom three rows are for aviation. In each of these experiments, performance was estimated first on a training device and later on a performance device.[86]

Table 8-7 breaks down the evaluation criteria based on the classes of variables developed in Chapter 7. All of these studies used two sets of Results measures to estimate transfer; that is, first during training and then post-training. (The Stewart study also collected Reaction data, but that is unrelated to the transfer question.) Table 8-7 provides a useful way to visualize what is meant by transfer. Transfer is estimated by comparing pairs of variables obtained during training with those obtained post-training. In these six cases, all of the variables represent comparable pairs of Results. Moreover, the data on both sets of variables were gathered in a relatively short period of time. Other types of transfer may also be of interest. It all depends upon the definition of Results. The narrowest way to define this term is as the same variable but collected under different circumstances.

[86] The common practice is to train on the simulator and test on operational equipment. However, as these evaluations demonstrate, other possible sequences are possible; for example, simulator A to simulator B, operational equipment to simulator.

Table 8-6. Descriptive Summary of Representative Evaluations for Evaluation Objective 3B: Measure Transfer of Training

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Witmer (1988)	Gunnery accuracy and speed (opening time, identification time, hit time, aiming error)	experiment (transfer)	Experiment: 2 groups of 12 M60A3 gunners trained on VIGS and UCOFT in opposite orders. Performance was assessed on second trainer used. Performance improvement (speed & accuracy) on each device was recorded and transfer from one device to the other was estimated by correlating scores between two devices.
Wheaton, Rose, Fingerman, Leonard, & Boycan (1976)	Percent hits, time between 1st. and 2nd. rounds, percent transfer	experiment (transfer)	4-group experiment: 3 groups trained on different burst on target training devices (17-4, 17-4 modified, COFT), control group practiced criterion task without prior training. All groups were then tested on M60A1 using laser firing device.
Bauer (1978)	Gunnery accuracy and speed	experiment (transfer)	3-group experiment. 2 groups practiced on mini range (130 or 260 rounds) and a third group used 7.62 coaxial MG preliminary tables. All groups then fired 105mm on Tables IV & VIII.
Kaempf & Blackwell (1990)	Performance on selected flight maneuvers	experiment (transfer)	2-group experiment: 20 aviators were pretested on flight skills during aircraft checkride and on simulator. 10 each were assigned to experimental and control groups. Experimental group trained to proficiency on simulator and then received similar training on aircraft. Control group trained to proficiency on aircraft and then was tested on simulator. Experimental group required little aircraft time to reach proficiency on aircraft; good transfer from simulator. Control group flying skills did not (backward) transfer to simulator.
Thorpe, Varney, McFadden, LeMaster, & Short (1978)	Student landing proficiency	experiment (transfer)	Three (3) groups of students received training in different flight simulators and were subsequently evaluated during flights and landings on KC-135.
Stewart (1994)	Flying performance	experiment (transfer) judgment (users)	Experienced pilots followed mission scenario on simulator and their performance was evaluated; participants also rated simulator flight characteristics.

Table 8-7. Evaluation Criteria by Study for Objective 3B: Measure Transfer of Training

WHEN	EVALUATION CRITERIA	STUDY					
		Witmer	Wheaton et al.	Bauer	Kaempf & Blackwell	Thorpe et al.	Stewart
During training	1. Reaction						√
	2. Collective Performance						
	3. Results	√	√	√	√	√	√
Post-training	4. Collective Performance						
	5. Results	√	√	√	√	√	√

Collective Performance Transfer

It is possible to compare collective performance during and post-training. Although none of the evaluations in Table 8-6 did this, it is not much of a stretch. For example, if any of the evaluations had focused on Collective Performance versus Results, transfer could be determined based on the Collective Performance variable during and post-training. To do this would require that data be collected and stored both during and after training in a form that would permit comparison.

Collective performance data of this type have seldom been collected or subjected to such analyses. This may be due to a traditional focus on Results variables in transfer studies, to the relative immaturity of collective performance assessment methods, because no one had given the idea much thought, or for other reasons. However, as the military makes increasing use of LSTS, with their emphasis on collective versus individual training, it makes sense to consider adding new variables to the set traditionally used in transfer studies.

Broadening the Definition of Transfer

Military training assumes that performance during training will affect job performance, which will affect combat readiness, which will affect combat performance (Solomon, 1986). Learning on a simulator should affect performance at various removes from the simulator to performance in field exercises, live simulations of combat (e.g., at the NTC), other simulations, and combat.

Recall Alluisi's argument about the importance of establishing the connection between training and unit readiness (see Chapter 7). The basis for establishing any possible linkage between simulator training and these more remote-from-training variables is to maintain archival data for units trained with and without the LSTS. Hiller (1994, 7 February), Boldovici and Bessemer (1994), and Leibrecht (1996) have all endorsed this strategy.[87] Once these data are available, they can be subjected to what Chapter 3 calls *ex post facto* analysis.[88] The legitimacy and appropriateness of what statisticians refer to by this name is controversial, to say the least. If the reader has doubts about applying such analyses, call the local statistician for help in deciding what to do next. The *ex post facto* approach has potential for evaluating LSTS in both the correlation/regression and comparison forms:

[87] As noted in Chapter 4, Hiller made the case that traditional experimental design could not be used to estimate the effects of CCTT on readiness and proposed a two aspect evaluation strategy: (1) long-term data collection from units training with/without SIMNET/CCTT, (2) separate, targeted experimental applications of CCTT.

[88] Recall from Chapter 3 that for purposes of this manual, *ex post facto* "experiments" were defined as studies that use historical data to mimic experiments. Fifteen studies in TCEF were classified as ex post facto based on their methodological descriptions. These appear to fall into two classes: ***comparison*** and ***correlation/regression***. Comparison studies, like 2- or more-group experiments, compare the effects of one or more experimental treatments, but based on historical rather than freshly-generated data. Correlation/regression studies use one of those statistical methods on historical data to calculate the degree to which a particular type of training contributes to later performance.

- Correlation/Regression: Here, the idea is to correlate selected during-training Results with post-training Results; for example, correlate historical data for performance in the simulator with performance in field exercises, in live simulations of combat, in other simulators, and with unit readiness. Bessemer (1998, 13 August) commented, "correlation does not establish cause, but it raises suspicions. Follow-on evaluations can use correlation methods to derive causal hypotheses for later test by quasi- or true experiments. This is a common TQM approach pursued to examine alternatives derived from cause-effect and flow-chart analyses of processes."
- Comparison: Archival data can be used to conduct comparison studies such as Derrick and Davis (1993). To do this, historical data must be collected over time for units trained with and without LSTS. An ex post facto "experiment" then compares data for the two conditions. Analyze enough data and it may be possible to estimate the effects on Reaction, Collective Performance, and Results. Granted, it would be preferable to conduct a traditional experiment. However, if this is not feasible, consider the approach described.

3C. Determine Transfer of Training

After completing system development, it is possible to determine whether positive transfer of training occurs. This determination could be made analytically by reviewing the literature relating to transfer in the training content area. No reviews of LSTS transfer have yet been published; the reviews cited under Sub-objective 2B deal with training effectiveness rather than transfer. Many reviews of transfer in the training content areas of flying and gunnery have been published and these may serve as surrogates to illustrate the concept:

- Morrison, Drucker, and Campshure (1991). Review of research on utility of devices and aids for training tank gunnery. Devices/aids covered are M1 TopGun, M1 Videodisc Interactive Gunnery Simulator (VIGS), M1 Mobile Conduct of Fire Trainer (M-COFT), Guard Unit Armory Device Full-Crew Interactive Simulation Trainer, Armor (GUARD FIST I), SIMNET, and hand held tutor. Comparative training effectiveness and transfer of devices was assessed. Evaluation criteria: gunnery performance. Method: analysis (evaluate).
- Orlansky and String (1977). Review of the effectiveness and costs of flight simulators. Findings: operating cost of flight simulators is estimated to be between 5-20% of the cost of aircraft. Many studies have shown that skills learned in flight simulators do transfer successfully to aircraft; the use of simulators for training can reduce flight time. Evaluation criteria: flying performance, degree of transfer of flying skills from simulator to aircraft, cost. Method: analysis (evaluate).

4. Predict, Measure, or Determine User Acceptance

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
4	Predict user acceptance	Measure user acceptance	Determine user acceptance

Evaluations may be conducted to predict, measure, or determine user acceptance.[89] Such estimates are based on judgment or survey. Evaluation criteria would be attitudes and judgments about training system effectiveness; for example, ratings of its training value on a scale from 1-10. Responses to open-ended questions and comments are other common ways to gather data to satisfy this objective.

4A. Predict User Acceptance

Predicting user acceptance is comparable to doing marketing research for a new product; for example, to estimate customer desire for the product. Objective 1 (Estimate need for a new training system) does this in part by establishing technical need. Still, a product may be needed but no one may want it. Presumably, before information about user preferences can be obtained, the product must exist in the form of a written description, prototype, or other tangible representation that potential users can consider and render judgments about. It is reasonable to survey users before proceeding to avoid developing a product that will later be rejected by customers. The survey might describe the product (e.g., training system, feature, attribute, innovation, etc.) and ask customers to estimate need, value, suggested alternatives, and so forth. TCEF contains no evaluations conducted to satisfy this objective.

4B. Measure User Acceptance

The top two rows in Table 8-8 provide information on two evaluations for Sub-objective 4B. Both of these were conducted on new LSTS that had been used by the respondents. The Fletcher evaluation was conducted early in SIMNET development. The Mirabella et al. evaluation was conducted at a comparable stage of MDT2 development. Both of these evaluations were based on user judgment.

[89] It is debatable whether this objective should be listed separately from 2A, B, C (predict, measure, determine training effectiveness) because it has already been recommended that Objective 2 obtain *Reaction* measures, which often include measures of user acceptance. It is listed separately here to underline the importance of obtaining this type of information and acknowledge that doing so is not the universal practice. Also, sometimes one is interested in user acceptance alone.

Table 8-8. Descriptive Summary of Representative Evaluations for Evaluation Objective 4B (Measure User Acceptance) and 4C (Determine User Acceptance)

OBJ	AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
4B	Mirabella, Sticha, & Morrison (1997)	Attitudes, judgments, and opinions re MDT2 simulation and its value	judgment (users)	User reactions to participation in MDT2 training were obtained with a combination of survey questionnaires, group interviews, and observations of training. Report discusses problems in conducting data collection and details questions asked, data collection procedures and instruments, and research findings. Includes lessons learned.
4B	Fletcher (1988)	Attitudes and judgments re SIMNET simulation and its value	judgment (users)	During early phase of SIMNET implementation commanders and crews at all levels were asked to rate and provide comments regarding the performance of SIMNET as a device and a simulator, how well it exercised different skills, its appropriate training role, and user acceptance.
4C	Sheppe, Sheppard, & McDonald (1990)	Attitudes and judgments re trainer effectiveness, utilization, and acceptance	judgment (users) judgment (SMEs)	Fleet personnel completed questionnaires and were interviewed to determine perceptions of effectiveness, utilization, and acceptance.
4C	Johnson (1995)	Supervisor/leader satisfaction with EO program and performance of DEOMI graduates	judgment (users) judgment (SMEs)	Senior leaders, commanders, and supervisors completed questionnaires and were interviewed to determine satisfaction with equal opportunity program and performance of graduates.

4C. Determine User Acceptance

The bottom two rows in Table 8-8 provide information on two evaluations for Sub-objective 4C. Neither of these is for an LSTS because no long-term evaluations of these simulations have yet been conducted. The Sheppe et al. evaluation is for the Navy's mobile pierside training program. The Johnson evaluation is for the DEOMI (Defense Equal Opportunity Management Institute) training program. Both of these evaluations were conducted after training was operational for long enough that users were able to judge its value. Both of these evaluations were based on user judgment.

5. Support Training Design; Determine Training Status

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
5	Support training design		Determine training status

During training development, evaluations may be conducted to support training design; for example, to select among alternative training strategies, to make tradeoffs among alternative training methods, to identify and correct training problems, and to otherwise aid training decision-making. Post-development, evaluations may be conducted to determine training status; for example, how well individuals in a particular MOS are able to perform their jobs. It is easy to see how these evaluations can both help influence training in the beginning and determine how well it is working after it becomes operational. Such studies provide feedback to assure that training, once fielded, is not forgotten. Such feedback is an essential element of TQM.

5AB. Support Training Design

Table 8-9 provides information for a sample of evaluations for Sub-objective 5AB. These evaluations were conducted during training development. All of the studies used analysis and three of the four also used judgment. Only the Keller et al. and Scott et al. evaluations deal with LSTS, although all of the studies concern simulators. Keller et al. analytically compared four alternative ways to train helicopter units. Berg et al. deals with the tradeoff between simulators and live fire for marksmanship training, Scott et al. with alternative ways to field the CCTT, and Rozen with the tradeoff between simulators and flying for maintaining flight proficiency.

Table 8-9. Descriptive Summary of Representative Evaluations for Evaluation Objective 5AB: Support Training Design

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Keller, Maruna, Hawkins, & Bealieu (1991)	Percent of trainable tasks on list	analysis (compare)	Analytical study to assess alternative ways to train helicopter units. Collective training tasks were identified and cost and effectiveness of four alternative ways to train was estimated: (1) aircraft without MILES, (2) aircraft with MILES/AGES, (3) simulator with AVCATT technology, (4) simulator with commercial technology. Alternatives were compared analytically. Alternative 3 offered best training capability.
Berg, Adedeji, & Steadman (1993)	Marksmanship accuracy, cost	analysis (optimize) judgment (SME)	An analytical and judgment-based study to estimate the extent to which the Marine Corps should use simulators vs. live-fire to perform infantry training tasks. Study applied a CNA-developed cost-effectiveness estimation method to gather effectiveness data from SMEs and combine it with CNA cost estimates to determine the appropriate mix of simulation and live fire. Effectiveness estimates and detailed cost data are provided. Study concluded that third-generation simulators can be used cost-effectively, that procuring them would be a very good investment, that they would increase the overall quality and effectiveness of training, and significantly reduce the total annual cost of training.
Scott, Djang, & Laferriere (1995)	Cost, training effectiveness	analysis (optimize) judgment (users)	Objective was to find best way to field future CCTT into reserves. Reserve soldiers with CCTT experience rated effectiveness of current training; ratings provided estimates of best training mission scenarios. Mathematical models were use to estimate costs of three fielding alternatives. Data collection instruments and mathematical models are described in detail
Rozen (1985)	Cost and training effectiveness	analysis (optimize) judgment (users)	Objective was to determine the most cost-effective combination of flight simulation and flying for maintaining proficiency of three categories of Israeli pilots (fighter, transport, helicopter). A cost effectiveness decision model is described. 58 pilots were interviewed and expressed their judgments on the best mix of simulators/flying. Data were used to construct transfer curves (isoquants) in accordance with Povenmire and Roscoe's method. Costs were estimated based on the curves generated. Sensitivity analyses were conducted. The method described is original and unique, combining cost analysis and linear modeling.

5C. Determine Training Status

Table 8-10 provides information for a sample of evaluations for Sub-objective 5C. These evaluations were conducted after training development. Most of the studies used experiments of the pre-experimental or test subtype to test users in what might be characterized as "competency tests;" that is, tests to determine whether personnel were able to perform their jobs up to certain predefined standards. The Ellis and Parchman study used a specialized analytical method to evaluate traditional and CBI-based versions of a course. George et al. used a survey to evaluate training.

Table 8-10. Descriptive Summary of Representative Evaluations for Evaluation Objective 5C: Determine Training Status

AUTHOR (YEAR)	EVALUATION CRITERIA	METHOD	SUMMARY
Ellis & Parchman (1994)	Student test performance and attitudes	analysis (compare) judgment (users)	The Course Evaluation System (CES) method was used to assess match between course objectives, test items, and instructional presentation for both new (CBI-based) and traditional versions of course. Students completed a questionnaire to assess attention, relevance, confidence, and satisfaction. Test scores were compared between new and old versions of course. No significant differences between old and new courses, but student questionnaire responses favored new course.
TEXCOM Combined Arms Test Center (1997)	Team performance on various tactical tasks	experiment (pre-) judgment (SME)	Various company level teams participated in training of their own choice, and later were tested for the record. Teams were then pretested, CCTT trained, and posttested. No actual performance data were recorded. Appear to have been many counfounding factors in test. Results are primarily observational.
Pishel, Neal, & Stapp (1991)	Performance test scores, survey, Interviews	experiment (test)	MCS operators/users were surveyed, interviewed, and performance tested
Wood (1987)	Ability to support training on nonsystem training device requirements	experiment (test)	Four battalions attempted to use software to support operation orders and to train staff. Numerous problems were encountered.
Salter (1998)	Performance on ARTEP tasks (against standard)	experiment (test) judgment (users)	Small-scale test of FIST-B training device. Bradley squads trained with device. Data were gathered during exercise re performance on tasks to be trained to pre-specified standard. At conclusion of training, squads were interviewed and completed questionnaires.
George, Jackson, Kenney, & Kilgore (1991)	STAMIS-TACCS operator test performance, attitudes of operators and managers/ supervisors	survey	Survey team visited several sites to conduct tests, surveys, and interviews. Objective was to assess training to support STAMIS-TACCS. STAMIS-TACCS operators were tested and surveyed, software analysts and managers/supervisors were surveyed and interviewed.

6. Support System Design; Evaluate System Design

CODE	PRE-DEVELOPMENT (A)	DEVELOPMENTAL (B)	POST-DEVELOPMENT (C)
6	Support system design		Evaluate system design

Need for System Design Studies

Studies should be conducted during system development to assure that the design provides an effective learning environment. After development is complete, the design should be further evaluated from that perspective. Few studies of this type have been published.[90] It is not clear if that is because such studies are rarely conducted or are conducted but rarely published. Whatever the case, training evaluators should participate in system design to assure that the systems provide an effective learning environment. Recall from Chapter 7 that milestone evaluations need only satisfy general requirements as specified in documents such as the ORD. Moreover, evaluators have considerable discretion in terms of what to evaluate and what to use as evaluation criteria. Program managers may choose to focus on hardware and software and not on training effectiveness. If this interest is to be represented, someone must advance it and act as its proponent.

In 1997, the Navy, lead Service for the JSIMS development, chartered the LMWG to influence JSIMS design from a learning perspective, acting in the proponent role mentioned in the previous paragraph. The LMWG documented its methods in *JSIMS Learning Methodology Reference Document: A Guide for System Designers and Developers* (Learning Methodology Working Group, 1999). While these methods were developed for JSIMS, they should apply to LSTS generally.

[90] Bessemer (1998, 13 August) commented that "simulator development could include user testing and experimental evaluation of alternative features and configurations if the acquisition system allowed the possibility, and requests-for-proposal were written to include that option." Historically, acquisition documents have not included this option and such testing has been the exception, rather than the rule.

Reference List 8-1. Evaluation Framework

Orlansky, J., Dahlman, C.J., Hammon, C.P., Metzko, J., Taylor, H.L., & Youngblut, C. (1994). *The value of simulation for training.* IDA Paper P-2982. Alexandria, VA: Institute for Defense Analyses. (ADA289174)

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33*(3), 343-362.

Morrison, J.E., Drucker, E.H., & Campshure, D.A. (1991). *Devices and aids for training M1 tank gunnery in the Army National Guard: A review of military documents and the research literature.* RR 1586. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA240628)

Orlansky, J. & String, J. (1977). *Cost-effectiveness of flight simulators for military training volume I: Use and effectiveness of flight simulators.* IDA Paper P-1275. Alexandria, VA: Institute for Defense Analyses. (ADA052801)

Ellis, J.A. & Parchman, S. (1994). *The interactive multisensor analysis training (IMAT) system: A formative evaluation in the aviation antisubmarine warfare operator (AW) class "A" school.* NPRDC TN-94-20. San Diego, CA: Navy Personnel Research and Development Center. (ADA285959)

TEXCOM (1997). *Test data report for the Close Combat Tactical Trainer limited user test.* TDR-97-LUT-1645A. Ft. Hood, TX: Author. (ADB228904)

Pishel, R.G., Neal, M.A., & Stapp, K.M. (1991). *Maneuver control system training effectiveness analysis.* TRAC-WSMR-TEA-91-005. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB158616)

Wood, M.A. (1987). *Final report on-site user test (OSUT) of Brigade/Battalion Battle Simulation (BBS).* Report No. USACSTA-6602. Aberdeen Proving Grounds, MD: U.S. Army Combat Systems Test Activity. (ADB117073)

Salter, M.S. (1998). *Full crew interactive simulation trainer -- Bradley (FIST-B): Limited user assessment.* RR 1724. Ft. Benning, GA: U.S. Army Research Institute Field Unit. (ADA345818)

George, E., Jackson, G., Kenney, M., & Kilgore, E. (1991).*Combat service support (CSS) standard Army management information systems (STAMIS) tactical Army combat service support computer system (TACCS) training effectiveness analysis (TEA) update.* White Sands Missile Range, NM: TRADOC Analysis Command. (ADB166701)

Ambruster, R.F. (1987). *Training effectiveness evaluation: OH-58D pilot and observer combat skills instruction.* Ft. Rucker, AL: U.S. Army Aviation Center. (ADA190938)

Bailey, S.S. & Hodak, G.W. (1994). *Live-fire versus simulation: A review of the literature.* Special Report 94-002. Orlando, FL: Naval Air Warfare Center Training Systems Division. (ADB189243)

Bauer, R.W. (1978). *Training transfer from mini-tank range to tank main gun firing.* Technical Paper 285. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Berg, R.M., Adedeji, A.M., & Steadman, G.W. (1993). *Simulation offset to live fire training phase 2 results: Application of the at least equal effectiveness methodology to simulator use in Marine Corps infantry training programs.* CRM-93-112. Alexandria, VA: Center for Naval Analyses. (ADB177151)

Berg, R.M., Adedeji, A.M., & Trenholm, C. (1993). *Simulation offset to live fire training study: Assessment of Marine Corps live fire training support.* CIM-238. Alexandria, VA: Center for Naval Analyses. (ADB173795)

Bessemer, D.W. (1998, 13 August). *Review of "proposed training effectiveness evaluation framework for large-scale training simulations (review draft).* Memorandum. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.

Boldovici, J.A. & Bessemer, D.W. (1994). *Training research with distributed interactive simulation: Lessons learned from simulation networking.* TR 1006. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285584)

Boldovici, J.A. (1987). Measuring transfer in military settings. In S.M. Cormier & J.D. Hagman (eds.), *Transfer of learning: Contemporary Research and Applications.* San Diego, CA: Academic Press.

Bretl, D.C., Rivera, F.G., & Coffey, B.A. (1996). *Engineer combined arms tactical trainer (ENCATT) cost and training effectiveness analysis (CTEA).* TRAC-WSMR-CTEA-96-008. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB213104)

Burnside, B.L. (1990). *Assessing the capabilities of training simulations: A method and simulation network (SIMNET) application.* ARI RR 1565. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226354)

Carroll, D.K. (1995). *Heavy assault bridge cost and training effectiveness analysis (CTEA) final report.* USAES-CTEA-95-HAB-001. Ft. Leonard Wood, MO: U.S. Army Engineer School. (ADA311901)

Derrick, D.L. & Davis, M.S. (1993). *Cost-effectiveness analysis of the C-130 aircrew training system.* AL-TR-1992-0173. Williams Air Force Base, AZ: Armstrong Laboratory. (ADB171592)

Drucker, E.H. & Campshure, D.A. (1990). *An analysis of tank platoon operations and their simulation on simulation networking (SIMNET).* RP 90-22. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA017009)

Evans, R.M. & Braby, R. (1983). *Self-paced and conventional instruction in Navy training: A comparison on elements of quality.* TAEG TR-147. Orlando, FL: Training Analysis and Evaluation Group. (ADA132402)

Finley, D.L. (1997). *Simulation-based communications realism and platoon training in the close combat tactical trainer (CCTT).* TR 1064. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337692)

Fletcher, J.D. (1988). *Responses of the 1/10 cavalry to SIMNET.* IDA Analysis Memorandum No. M-494. Arlington, VA: Defense Sciences Office. (ADA200499).

Hiller, J.W. (1994, 7 February). *Close combat tactical trainer (CCTT) evaluation planning.* Memorandum to COL Thomas A. Horton. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Johnson, J.L. (1995). *A preliminary investigation into DEOMI training effectiveness.* RSP 95-8. Patrick Air Force Base, FL: Defense Equal Opportunity Management Institute. (ADA300958)

Kaempf, G.L. & Blackwell, N.J. (1990). *Transfer of training study of emergency touchdown maneuvers in the AH-1 flight and weapons simulator.* RR 1561. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226360)

Keller, A.R., Maruna, R.E., Hawkins, K.A., & Bealieu, H.H. (1991). *Aviation combined arms tactical training development study (TDS) – Phase II.* Report No. AVNC-DOTD-92-1. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB161932)

Kelly, J.F. (1995). *A training effectiveness evaluation of leathernet.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA303556)

Kraemer, R.E. & Rowatt, W.C. (1993). *A review and annotated bibliography of armor gunnery training device effectiveness literature.* RR 1652. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275258)

Learning Methodology Working Group (1999). *JSIMS learning methodology reference document: A guide for system designers and developers.* LMWG Reference Guide 99-001. Orlando, FL: Author.

Leatherwood, N.J., Schisser, J.S. & Russell, R.J. (1986). *OH-58D aircrew cost and training effectiveness analysis (OH-58D aircrew CTEA): Final report.* TRADOC ACN 85216. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB112520)

Leibrecht, B.C. (1996). *An integrated database of unit training performance: Description and lessons learned.* Orlando, FL: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328669)

Lickteig, C.W. & Collins, J.W. (1995). *Combat vehicle command and control system evaluation: Vertical integration of an armor battalion.* TR 1021. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292718)

Lynn, J. & Palmer, K.L. (1991). *Independent operational assessment of the close combat tactical trainer (CCTT).* OA-0200. Alexandria, VA: U.S. Army Operational Test and Evaluation Command. (ADB160088)

McDade, M.B. (1986). *Bradley fighting vehicle (BFV) training developments study (TDS).* Ft. Benning, GA: Analysis and Studies Office, U.S. Army Infantry School. (ADA173795)

Mirabella, A., Sticha, P., & Morrison, J. (1997). *Assessment of user reactions to the multi-service distributed training testbed (MDT2) system.* TR 1061. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328473)

Noble, J.L. & Johnson D.R. (1991a). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 1: Executive summary.* TRAC-WSMR-CTEA-91-018-1. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157064)

Noble, J.L. & Johnson D.R. (1991b). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 2: Main report.* TRAC-WSMR-CTEA-91-018-2. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB173567)

Orlansky, J., Taylor, H.L., Levine, D.B., & Honig, J.G. (1997). *The cost and effectiveness of the multi-service distributed training testbed (MDT2) for training close air support.* IDA Paper P-3284. Alexandria, VA: Institute for Defense Analyses.

Rozen, U. (1985). *Analyzing the cost effectiveness of using flight simulators in the Israeli air force.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA164864)

Scott, B.B., Djang, P., Laferriere, R. (1995). *Reserve component (RC) mobile close combat tactical trainer (M-CCTT) integration and deployment study.* TRAC-WSMR-TR-95-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Center. (ADB202913)

Sheppe, M.L., Sheppard, D.J., & McDonald, R.G. (1990). *Waterfront trainer program evaluation of phase II.* TR 90-017. Orlando, FL: Naval Training Systems Center. (ADB151857)

Shlechter, T.M., Bessemer, D.W., Nesselroade, P., & Anthony, J. (1995). *An initial evaluation of a simulation-based training program for Army National Guard units.* ARI RR 1679. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA297271)

Shlechter, T.M., Kraemer, R.E., Bessemer, D.W., Burnside, B.L., & Anthony, J. (1996). *Perspectives on the virtual training program from members of its initial observer/controller team.* ARI-RR-1691. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA310080)

Smith, B.W., & Cross, K.D. (1992). *Assessment of Army Aviators' Ability to Perform Individual and Collective Tasks in the Aviation Networked Simulator (AIRNET).* RN-92-32. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA250293)

Solomon, H. (1986). *Economic issues in cost-effectiveness analyses of military skill training.* IDA Paper P-1897. Alexandria, VA: Institute for Defense Analyses. (AD-A171106)

Stewart, J.E. (1994). *Using the backward transfer paradigm to validate the AH-64 simulator training research advanced testbed for aviation.* RR 1666. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285758)

Thomas, B.W. & Gainer, C.A. (1990, May). Simulation networking: Low fidelity simulation in U.S. Army aviation. *Proceedings of the Royal Aeronautical Society* (pp. 18.1-18.11). London, England.

Thorpe, J.A., Varney, N.C., McFadden, R.W., LeMaster, W.D., & Short, L.H. (1978). *Training effectiveness of three types of visual systems for KC-135 flight simulators.* AFHRL-TR-78-16. Brooks AFB, TX: Air Force Human Resources Laboratory. (ADA060253)

Universal Cities Studios, Inc. (1985). *Back to the future.* Motion Picture. Hollywood, CA: Author.

Wetzel, C. D., Simpson, H., & Seymour, G.E. (1995). *The use of videoteletraining to deliver chief and leading petty officer navy leadership training: evaluation and summary.* NPRDC-TR-95-8. San Diego, CA: Navy Personnel Research and Development Center. (ADA298374)

Wheaton, G. R., Rose, A. M., Fingerman, F.W., Leonard, R. L, and Boycan, G. G. (1976). *Evaluation of three burst-on-target trainers.* RM 76-18). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA076820)

Witmer, B.G. (1988). *Device-based gunnery training and transfer between the videodisk gunnery simulator (VIGS) and the unit conduct of fire trainer (U-COFT).* TR 794. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA197769)

Worley, R.D., Simpson, H.K., Moses, F.L., Aylward, M., Bailey, M., & Fish, D. (1996). *Utility of modeling and simulation in the department of defense: Initial data collection.* IDA Document D-1825. Alexandria, VA: Institute for Defense Analyses. (ADA312153)

REFERENCES

Alluisi, E.A. (1991). The development of technology for collective training: SIMNET, a case history. *Human Factors, 33*(3), 343-362.

Ambruster, R.F. (1987). *Training effectiveness evaluation: OH-58D pilot and observer combat skills instruction.* Ft. Rucker, AL: U.S. Army Aviation Center. (ADA190938)

Angier, B.N., Alluisi, E.A., Horowitz, S.A. (1992) *Simulators and Enhanced Training.* IDA Paper P-2672. Institute for Defense Analyses, Alexandria, VA, 1992.

Babbitt, B.A. & Nystrom, C.O. (1989a). *Questionnaire construction manual.* RP 89-20. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA212365)

Babbitt, B.A. & Nystrom, C.O. (1989b). *Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography.* RP 89-21. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA213255)

Bailey, S.S. & Hodak, G.W. (1994). *Live-fire versus simulation: A review of the literature.* Special Report 94-002. Orlando, FL: Naval Air Warfare Center Training Systems Division. (ADB189243)

Bauer, R.W. (1978). *Training transfer from mini-tank range to tank main gun firing.* Technical Paper 285. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bell, H. (1995). Symposium on distributed simulation for military training of teams/groups: The engineering of a training network. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting,* San Diego, CA, 1311-1315.

Bell, H. (1996). Panel of Multi-Service Distributed Training Testbed: DIS training of military teams/groups: The engineering of a training network. *Proceedings of the 17th. I/ITSEC Conference.* Albuquerque, NM, 365-370.

Bell, H.H. & Waag, W.L. (1998). Evaluating the effectiveness of flight simulators for training combat skills: A review. *International Journal of Aviation Psychology, 8*, 224-242.

Bell, H.H., Dwyer, D.J., Love, J.F., Meliza, L.L., Mirabella, & Moses, F.L. (1997a). *Recommendations for planning and conducting multi-service tactical training with distributed interactive simulation technology.* RP 97-03. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328480)

Bell, H.H., Dwyer, D.J., Love, J.F., Meliza, L.L., Mirabella, & Moses, F.L. (1997b). *Recommendations for planning and conducting multi-service tactical training with distributed interactive simulation technology: Appendices.* RP 97-04. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA336275)

Berg, R.M., Adedeji, A.M., & Steadman, G.W. (1993). *Simulation offset to live fire training phase 2 results: Application of the at least equal effectiveness methodology to simulator use in Marine Corps infantry training programs.* CRM-93-112. Alexandria, VA: Center for Naval Analyses. (ADB177151)

Berg, R.M., Adedeji, A.M., & Trenholm, C. (1993). *Simulation offset to live fire training study: Assessment of Marine Corps live fire training support.* CIM-238. Alexandria, VA: Center for Naval Analyses. (ADB173795)

Bessemer, D.W. (1991). *Transfer of SIMNET training in the armor officer basic course.* TR 920. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA233198)

Bessemer, D.W. (1998, 13 August). *Review of "proposed training effectiveness evaluation framework for large-scale training simulations (review draft).* Memorandum. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.

Bessemer, D.W. & Myers, W.E. (1998). *Sustaining and improving structured simulation-based training.* RR 1722. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA344895)

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Boldovici, J.A. (1987). Measuring transfer in military settings. In S.M. Cormier & J.D. Hagman (eds.), *Transfer of learning: Contemporary Research and Applications.* San Diego, CA: Academic Press.

Boldovici, J.A. & Bessemer, D.W. (1994). *Training research with distributed interactive simulation: Lessons learned from simulation networking.* TR 1006. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285584)

Boldovici, J.A. & Bessemer, D.W. (1999). *The elements of training evaluation.* Special Report. Alexandria, VA: U.S. Army Research Institute for Behavioral and Social Sciences.

Boldovici, J.A. & Kolasinski, E.M. (1997). How to make decisions about the effectiveness of device-based training: Elaborations on what everybody knows. *Military Psychology, 9,* 121-135.

Bouchard, T.J. (1976). Field research methods: Interviewing, questionnaires, participant observation, systematic observation, unobtrusive measures. In M.D. Dunnette (Ed.) *Handbook of Industrial and Organizational Psychology.* Chicago, IL: Rand McNally.

Bretl, D.C., Rivera, F.G., & Coffey, B.A. (1996). *Engineer combined arms tactical trainer (ENCATT) cost and training effectiveness analysis (CTEA).* TRAC-WSMR-CTEA-96-008. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB213104)

Brown, F.J. (1978a). *The Army training study: Administration.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186321)

Brown, F.J. (1978b). *The Army training study: Training effectiveness analysis, volume I: Armor.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186323)

Brown, F.J. (1978c). *The Army training study: Training effectiveness analysis, volume II: Armor.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186324)

Brown, F.J. (1978d). *The Army training study: Training effectiveness analysis, volume IV: Ordnance, signal, and computer assisted map maneuver system (CAMMS).* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186326)

Brown, F.J. (1978e). *The Army training study: Training effectiveness analysis: Summary.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186322)

Brown, F.J. (1978f). *The Army training study: Data book.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA184393)

Brown, F.J. (1978g). *The Army training study: Report summary.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA184392)

Brown, R. & Mullis, C. (1988a). *Simulation networking assessment of perceptions - I.* TRASANA Report No. LR-1-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB118627)

Brown, R. & Mullis, C. (1988b). *Simulation networking assessment of perceptions - II.* TRASANA Report No. LR-2-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB146645)

Brown, R.E., Pishel, R.G., & Southard L.D. (1988). *Simulation networking (SIMNET) preliminary training developments study.* TRAC-WSMR-TEA-8-99. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB120874)

Browning, R.F., McDaniel, W.C., Scott, P.G., & Smode, A.F. (1982). *An assessment of the training effectiveness of device 2F64C for training helicopter replacement pilots.* TR 127. Orlando, FL: Training Analysis and Evaluation Group. (ADA118942)

Burnside, B.L. (1990). *Assessing the capabilities of training simulations: A method and simulation network (SIMNET) application.* ARI RR 1565. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226354)

Campbell, C.H., Campbell, R.C., Sanders, J.J., Flynn, M.R., & Myers, W.E. (1995). *Methodology for the development of structured simulation-based training.* ARI RP 95-08. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA296171)

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D.T. & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* In N.L. Gage (Ed.). *Handbook of Research on Teaching.* Chicago, IL: Rand-McNally.

Campbell, D.T. & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Cannon-Bowers, J.A. & Salas, E. (1997). A framework for developing team performance measures in training. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team Performance Assessment and Measurement: Theory, Methods and Applications.* Hillsdale, NJ: Lawrence Erlbaum.

Carroll, D.K. (1995). *Heavy assault bridge cost and training effectiveness analysis (CTEA) final report.* USAES-CTEA-95-HAB-001. Ft. Leonard Wood, MO: U.S. Army Engineer School. (ADA311901)

Clapper, D. & Schwab, J. (1986). *Concept evaluation program of simulation networking (SIMNET).* Ft. Knox, KY: U.S. Army Armor and Engineering Board. (ADB114371)

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Colburn, E. Farrow, S., & McDonough, J. (1994). *ADST multi-service distributed training testbed (MDT2) lessons learned.* ADST/WDL/TR-94-W003312. Orlando, FL: Loral Systems ADST Program Office. (ADA282380)

Communications Technology Applications, Inc. (1988). *Joint tactical information distribution system (JTIDS) cost and training effectiveness analysis (CTEA).* McLean, VA: Author. (ADA221316)

Cook, T.D. & Campbell, D.T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. In M.D. Dunnette (Ed.). *Handbook of Industrial and Organizational Psychology* (pp. 223-326). Chicago, IL: Rand McNally.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago, IL: Rand McNally.

Cosby, N.L. (1995). *SIMNET: An insider's perspective.* IDA Document D1661. Alexandria, VA: Institute for Defense Analyses. (ADA294786)

Crane, P.M. & Berger, S.C. (1993). *Multiplayer simulator based training for air combat.* Williams AFB, AZ: Air Force Armstrong Laboratory.

Crawford, A. & Suchan, J. (1996). *Understanding videoteleducation: An overview.* NPS-SM-96-003. Monterey, CA: U.S. Naval Postgraduate School. (ADA319585)

Department of Defense (1990). *MIL-STD-1379D: Military training programs.* Washington, D.C.: Author.

Department of Defense (1995). *DoD instruction 5000.61: DoD modeling and simulation (M&S) verification, validation, and accreditation (VV&A).* Washington, DC: Author.

Department of Defense (1996a). *DoD directive 5000.1: Defense acquisition.* Washington, D.C.: Author. (ADM000591)

Department of Defense (1996b). *DoD regulation 5000.2-R: Mandatory procedures for major defense acquisition programs (MDAPs) and major automated information system (MAIS) acquisition programs.* Washington, D.C.: Author.

Department of Defense (1996c). *Verification, validation, and accreditation (VV&A) recommended practices guide.* Alexandria, VA: Defense Modeling and Simulation Office. (http://www.dmso.mil/docslib/mspolicy/vva/rpg/)

Department of Defense (1998). *DoD directive 5000.59-M. Glossary of modeling and simulation (M&S) terms.* Arlington, VA: Defense Modeling and Simulation Office.

Department of Defense, Office of the Inspector General (1997). *Requirements planning and impact on readiness of training simulators and devices.* Report No. 97-138. Arlington, VA: Author.

Department of the Army (1994). *TRADOC regulation 350-32: The TRADOC training effectiveness analysis (TEA) system.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command.

Derrick, D.L. & Davis, M.S. (1993). *Cost-effectiveness analysis of the C-130 aircrew training system.* AL-TR-1992-0173. Williams Air Force Base, AZ: Armstrong Laboratory. (ADB171592)

Djang, P.A., Butler, W.A., Laferriere, R.R., & Hughes, C.R. (1993). *Training mix model.* TRAC-WSMR-TEA-93-035. White Sands Missile Range, NM: TRADOC Analysis Center. (ADB178428)

Drucker, E.H. & Campshure, D.A. (1990). *An analysis of tank platoon operations and their simulation on simulation networking (SIMNET).* RP 90-22. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA017009)

Dwyer, D.J., Fowlkes, J., Oser, R.L., & Salas, E. (1996). Panel on multi-service distributed training testbed, DIS training of military teams/groups: Case study results using distributed interactive simulation for close air support. *Proceedings of the 1996 International Training Equipment Conference.* The Hague, Netherlands, 371-380.

Dwyer, D.J., Fowlkes, J.E., Oser, R.L., Salas, E., & Lane, N.E. (1997). Team performance measurement in distributed environments: the TARGETs methodology. In M.T. Brannick, E. Salas, & C. Prince (Eds.), *Team Performance Assessment and Measurement: Theory, Methods and Applications* (pp. 137-153). Hillsdale, NJ: Lawrence Erlbaum.

Dwyer, D.J., Oser, R.L., & Fowlkes, J.E. (1995). Symposium on distributed simulation for military training of teams/groups: A case study of distributed training and training performance. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, 1316-1320.

Ellis, J.A. & Parchman, S. (1994). *The interactive multisensor analysis training (IMAT) system: A formative evaluation in the aviation antisubmarine warfare operator (AW) class "A" school.* NPRDC TN-94-20. San Diego, CA: Navy Personnel Research and Development Center. (ADA285959)

Ennis, J.J. & Gardner, C.V. (1990). *Chaparral/FLIR post fielding training effectiveness analysis (PFTEA).* FB TEA 1-89. Ft.Bliss, TX: U.S. Army Air Defense Artillery School. (ADB156909)

Evans, R.M. & Braby, R. (1983). *Self-paced and conventional instruction in Navy training: A comparison on elements of quality.* TAEG TR-147. Orlando, FL: Training Analysis and Evaluation Group. (ADA132402)

Finley, D.L. (1997). *Simulation-based communications realism and platoon training in the close combat tactical trainer (CCTT).* TR 1064. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337692)

Fletcher, J.D. (1988). *Responses of the 1/10 cavalry to SIMNET.* IDA Analysis Memorandum No. M-494. Arlington, VA: Defense Sciences Office. (ADA200499).

Fober, G.W., Dyer, J.L., & Salter, M.S. (1994). Measurement of performance at the joint training center: Tools of assessment. In. R.F. Holz, J.H. Hiller, & H.H. McFann (Eds.) (1994). *Determinants of effective unit performance: Research on measuring and managing unit training readiness* (pp. 39-70). Book. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Fowler, F.J. (1993). *Survey research methods.* Beverly Hills, CA: Sage Publications

Fowlkes, J., Dwyer, D.J., Oser, R.L., & Salas, E. (1997). Event-based approach to training. *Proceedings of the 19th. I/ITSEC Conference.* Albuquerque, NM.

Fowlkes, J.E., Lane, N.E., Salas, E., Franz, T., & Oser, R. (1994). Improving the measurement of team performance: The TARGETs methodology. *Military Psychology, 6*, 47-61.

Fusha, J.E. (1989). *Simulation networking (SIMNET): Evaluation of institutional/USAIS (U. S. Army Infantry School) use of SIMNET-T. Phases 1 and 2.* RN 2-89. Ft. Benning, GA: U.S. Army Infantry School. (ADA137722)

Gage, N.L. (Ed.) (1963). *Handbook of research on teaching.* Chicago, IL: Rand-McNally.

Garlinger, D.K. & Fallesen, J.J. (1988). *Review of command group training measurement methods.* TR 798. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA201753)

George, E., Jackson, G., Kenney, M., & Kilgore, E. (1991).*Combat service support (CSS) standard Army management information systems (STAMIS) tactical Army combat service support computer system (TACCS) training effectiveness analysis (TEA) update.* White Sands Missile Range, NM: TRADOC Analysis Command. (ADB166701)

Glaser, R. (1976). Components of a psychology of instruction: Towards a science of design. *Review of Educational Research, 46*(1), 1-24.

Greene, D.E. & Haynes, M.D. (1988). *Concept evaluation program (CEP) test of the tube-launched, optically tracked, wire-guided (TOW) training strategy.* USAIB Project No. 3901. Ft. Benning, GA: U.S. Army Infantry Center. (ADB125492)

Gurwitz, R., Burke, E., Calvin, J., Chatterjee, A., & Harris, M. (1983). *Large-scale simulation network design study.* Bolt, Beraneck, & Newman. (ADA134662)

Hall, E.R, Rankin, W.C., & Aagard, J.A. (1976) *Training effectiveness assessment, volume II: Problems, concepts and evaluation alternatives.* TAEG Report 39. Orlando, FL: Training Analysis and Evaluation Group. (ADA036518)

Hall, E.R. & Rizzo, W.A. (1975). *An assessment of U.S. Navy tactical team training.* TR 18. Orlando, FL: Training Analysis and Evaluation Group. (ADA011452)

Harman, J. (1984). *Three years of evaluation of the Army's basic skills education program.* RR 1380. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA170476)

Harman, J., Bell, S.A., & Laughy, N. (1989). *Evaluation of the hand-held mathematics tutor.* RR 1509. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA207157)

Harris, K. (1996). *Battle force tactical training (BFTT) developmental test IIB (DT-IIB) developmental test report.* Unpublished test report. Port Hueneme, CA: Naval Surface Warfare Center.

Hart, R.J., Hagman, J.D., & Bowne, D.S. (1990). *Tank gunnery: Transfer of training from TOPGUN to the conduct-of-fire trainer.* RR 1560. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA223165)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990a). *Verification and Validation of SIMNET-T. Phase 1. K/DSRD-116.* Oak Ridge, TN: Martin Marietta Energy Systems, Inc. (ADB147354)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990b). *Verification and validation of SIMNET-T from ORDGRP (K/DSRD-117).* Ft. Leavenworth, KS: U.S. Army Training and Doctrine Command Analysis Command. (ADB147355)

Henerson, M.E., Morris, L.L., & Fitz-Gibbon, C.T. (1987). *How to measure attitudes.* (Volume 6 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Herman, J.L., Morris, L.L., & Fitz-Gibbon, C.T. (1987). *Evaluator's handbook.* (Volume 1 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Hiller, J.W. (1994, 7 February). *Close combat tactical trainer (CCTT) evaluation planning.* Memorandum to COL Thomas A. Horton. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hiller, J.H. (1997). *Successfully evaluating training devices in an imperfect world.* Unpublished manuscript. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hiller, J.W. (1998, 6 August). *Personal communication.*

Hoffman, G.R. (1997). C*ombat support and combat service support expansion to the virtual training program SIMNET battalion exercise: History and lessons learned.* RR 1717. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA341201)

Holstead, J.R. (1989). *Large scale simulation networking (SIMNET) operational effectiveness appraisal (OEA) final report.* TAC Project 89-190T. Eglin AFB, FL: U.S.A.F. Tactical Air Warfare Center. (ADB133378)

Holz, R.F., Hiller, J.H., & McFann, H.H. (Eds.) (1994). *Determinants of effective unit performance: Research on measuring and managing unit training readiness.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Jeantheau, G.G. (1971). *Handbook for training systems evaluation.* NAVEDTRACEN 66-C-0113-2. Darien, CT: Dunlap & Associates. (AD733962)

Johnson, E.M. & Baker, J.D. (1974). Field testing: The delicate compromise. *Human Factors, 16,* 203-214.

Johnson, J.L. (1995). *A preliminary investigation into DEOMI training effectiveness.* RSP 95-8. Patrick Air Force Base, FL: Defense Equal Opportunity Management Institute. (ADA300958)

Kaempf, G.L. (1986). Backward transfer of emergency touchdown maneuvers. In K.D. Cross & S.M. Szabo (Eds.), *Human factors research in aircrew performance and training: Final summary report.* RN 86-94, pp. 42-51. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA175348)

Kaempf, G.L. & Blackwell, N.J. (1990). *Transfer of training study of emergency touchdown maneuvers in the AH-1 flight and weapons simulator.* RR 1561. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226360)

Kaminski, P.G. (1997, March 17). *Memorandum for office of the inspector general, DoD (OIG, DoD). Subject: Response to OIG, DoD, draft audit report, "Requirements Planning for Development, Test, Evaluation, and Impact on Readiness of Training Simulators and Devices,"* Project No. 5AB-0070.00, January 10, 1997. Memorandum. Washington, DC: Under Secretary of Defense for Acquisition and Technology.

Kass, R., (1997). *Test officer's guide for designing valid tests and experiments* (job aid). Ft. Hood, TX: TEXCOM Combined Arms Test Center.

Kass, R., (1997, June/July). Design of valid operational tests. *ITEA Journal.*

Keller, A.R., Maruna, R.E., Hawkins, K.A., & Bealieu, H.H. (1991). *Aviation combined arms tactical training development study (TDS) – Phase II.* Report No. AVNC-DOTD-92-1. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB161932)

Keller, A.R., Parrish, J.J., Harrison, J.A., & Macklin, L. (1992). *Mobile aircrew sustainment training-Apache (MAST-A) training effectiveness analysis - Phase I (TEA-I).* Report No. AVNC-DOS-92-5. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB172118)

Kelly, J.F. (1995). *A training effectiveness evaluation of leathernet.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA303556)

Kirkpatrick, D.L. (1976). Evaluation of training. In Craig, R.L. (Ed.) (1976). *Training and development handbook: A guide to human resource development* (pp18-1-18-27). New York, NY: McGraw-Hill.

Klein, G.A., Johns, P., Perez, R., & Mirabella, A. (1985). *Comparison-based prediction of cost and effectiveness of training devices: A guidebook.* RP 85-29. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA170941)

Kraemer, R.E. & Bessemer, D.W. (1987). *U.S. tank platoon training for the 1987 Canadian army trophy (CAT) competition using a simulation networking (SIMNET) system.* RR 1457. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA191076)

Kraemer, R.E. & Rowatt, W.C. (1993). *A review and annotated bibliography of armor gunnery training device effectiveness literature.* RR 1652. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275258)

Kribs, H.D. & Mark, L.J. (1982). *An evaluation of media selection.* NPRDC Special Report 82-13. San Diego, CA: Navy Personnel Research and Development Center.

Kulik, C.C., Schwalb, B.J., & Kulik, J.A. (1982). Programmed instruction in secondary education: A meta-analysis of evaluation findings. *Journal of Educational Research, 75*(3), 133-138.

Kulik, C.-L.C., Kulik, J.A., & Bangert-Drowns, R.L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*, 265-299.

Kulik, J.A & Kulik, C.-L.C. (1984). Effects of accelerated instruction on students. *Review of Educational Research, 54*, 409-426.

Kulik, J.A., Kulik, C.C., & Cohen, P.A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34*(4), 307-318.

Lampton, D.R. (1989). *Evaluation of a low fidelity battle simulation for training and evaluating command, control, and communications (C3I) skills for the armor platoon leader.* RR 1521. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA210606)

Learning Methodology Working Group (1999). *JSIMS learning methodology reference document: A guide for system designers and developers.* LMWG Reference Guide 99-001. Orlando, FL: Author.

Leatherwood, N.J., Schisser, J.S. & Russell, R.J. (1986). *OH-58D aircrew cost and training effectiveness analysis (OH-58D aircrew CTEA): Final report.* TRADOC ACN 85216. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB112520)

Leibrecht, B.C. (1996). *An integrated database of unit training performance: Description and lessons learned.* Orlando, FL: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328669)

Lewman, T., Mullen, W.J., & Root, J. (1994). A conceptual framework for measuring unit performance. In. R.F. Holz, J.H. Hiller, & H.H. McFann (Eds.). *Determinants of effective unit performance: Research on measuring and managing unit training readiness* (pp. 17-38). Book. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292342)

Lickteig, C.W. & Collins, J.W. (1995). *Combat vehicle command and control system evaluation: Vertical integration of an armor battalion.* TR 1021. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292718)

Litwin, M.S. (1995). *How to measure survey reliability and validity.* Beverly Hills, CA: Sage Publications.

Loral Systems (1994). *Protocol extensions to DIS and interface requirements specification (IRS) for the multi-service distributed training testbed (MDT2).* Revision 1.0. ADST/WDL/TR--94-W003412. Orlando, FL: Author. (ADB198789)

Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *AERA Journal, 17*, 211-218.

Lynn, J. & Palmer, K.L. (1991). *Independent operational assessment of the close combat tactical trainer (CCTT).* OA-0200. Alexandria, VA: U.S. Army Operational Test and Evaluation Command. (ADB160088)

Lysakowski, R.S., & Walberg, H.J. (1981). Classroom reinforcement: A quantitative synthesis. *Journal of Educational Research, 75*, 39-77.

Lysakowski, R.S., & Walberg, H.J. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *AERA Journal, 19*, 559-578.

McDade, M.B. (1986). *Bradley fighting vehicle (BFV) training developments study (TDS).* Ft. Benning, GA: Analysis and Studies Office, U.S. Army Infantry School. (ADA173795)

McDaniel, W.C. (1987). *Training effectiveness evaluation of aviation antisubmarine warfare basic operator trainer (Device 14D1).* TR 87-019. Orlando, FL: Naval Training Systems Center. (ADB118742)

Meister, D. (1986). *Human factors testing and evaluation.* New York: Elsevier.

Meister, D. & Rabideau, G.F. (1965). *Human factors evaluation in system development.* New York: Wiley.

Meliza, L.L. (1993). *Simulation networking/training requirements relational database: User's guide.* RP 94-01. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA275634)

Meliza, L.L., Bessemer, D.W., & Tan, S.C. (1992). *Unit performance assessment systems development.* TR 1008. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285805)

Meliza, L.L., Bessemer, D.W., Burnside, B.L., & Shlechter, T.M. (1992). *Platoon-level after action review aids in the SIMNET unit performance assessment system (UPAS).* TR 956. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA254909)

Meliza, L.L. & Tan, S.C. (1992). *SIMNET unit performance assessment system (UPAS) version 2.5 user's guide.* ARI-RP-96-05. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA318046)

Merriam-Webster (1986). *Webster's ninth new collegiate dictionary.* Springfield, MA: Merriam-Webster. Author.

Mirabella, A. (1995). Symposium on distributed simulation for military training of teams/groups: MDT2 system assessment and effectiveness. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting*, San Diego, CA, 1321-1325.

Mirabella, A. (1998, July 31). *Personal communication.*

Mirabella, A., Sticha, P., & Morrison, J. (1997). *Assessment of user reactions to the multi-service distributed training testbed (MDT2) system.* ARI TR 1061. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328473)

Morrison, J.E. & Hoffman, R.G. (1992). *A user's introduction to determining tradeoffs among tank gunnery training methods.* ARI RN 92-29. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA250029)

Morrison, J.E., Drucker, E.H., & Campshure, D.A. (1991). *Devices and aids for training M1 tank gunnery in the Army National Guard: A review of military documents and the research literature.* RR 1586. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA240628)

Moses, F. L. (1995). Symposium on distributed simulation for military training of teams/groups: The challenge of distributed training. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting,* San Diego, CA, 1306-1310.

Muckler, F.A. & Finley, D.L. (1994a). *Applying training system estimation models to army training, Volume I: Analysis of the literature.* ARL-TR-463. Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory. (ADA283021)

Muckler, F.A. & Finley, D.L. (1994b). *Applying training system estimation models to army training, Volume II: An annotated bibliography 1970-1990.* ARL-TR-463. Aberdeen Proving Grounds, MD: U.S. Army Research Laboratory. (ADA283022)

Muller, D., Adkins, S., Belfer, B., Carter, J., & Levy, L. (1988). *The non line of-sight weapon system cost and training effectiveness analysis.* Ft. Bliss, TX: U.S. Army Air Defense Artillery School. (ADB147262)

Noble, J.L. & Johnson D.R. (1991a). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 1: Executive summary.* TRAC-WSMR-CTEA-91-018-1. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157064)

Noble, J.L. & Johnson D.R. (1991b). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 2: Main report.* TRAC-WSMR-CTEA-91-018-2. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB173567)

Orlansky, J. (1985). *Panel on the defence applications of operational research.* DA/A/DR (85)167. Proceedings of the Symposium on the Military Value and Cost-Effectiveness of Training, 7-10 January 1985. NATO, Brussels, Belgium. (AD-B-093-505)

Orlansky, J. (1989). *Executive Summary.* AC/243 (Panel 7/RSG.15)D/4 on the Military Value and Cost-Effectiveness of Training. NATO, Brussels, Belgium.

Orlansky, J., Dahlman, C.J., Hammon, C.P., Metzko, J., Taylor, H.L., & Youngblut, C. (1994). *The value of simulation for training.* IDA Paper P-2982. Alexandria, VA: Institute for Defense Analyses. (ADA289174)

Orlansky, J. & String, J. (1977). *Cost-effectiveness of flight simulators for military training volume I: Use and effectiveness of flight simulators.* IDA Paper P-1275. Alexandria, VA: Institute for Defense Analyses. (ADA052801)

Orlansky, J., Taylor, H.L., Levine, D.B., & Honig, J.G. (1997). *The cost and effectiveness of the multi-service distributed training testbed (MDT2) for training close air support.* IDA Paper P-3284. Alexandria, VA: Institute for Defense Analyses.

Paschal, R., Weinstein, T., & Walberg, H.J. (1984). The effects of homework on learning: A quantitative synthesis. *Journal of Educational Research, 78*(2), 97-104.

Patton, M.Q. (1987). *How to use qualitative methods in evaluation.* (Volume 4 of Sage Program Evaluation Kit.) Beverly Hills, CA: Sage Publications.

Pfeiffer, M.G. & Browning, R.F. (1984). *Field evaluations of aviation trainers.* NTSC TR-157. Orlando, FL: Naval Training Systems Center. (ADB083584).

Pfeiffer, M.G., Evans, R.M, & Ford, L.H. (1985). *Modeling field evaluations of aviation trainers.* TAEG TN 1-85. Orlando, FL: Training Analysis and Evaluation Group.

Pfeiffer, M.G. & Horey, J.D. (1988). *Analytic approaches to forecasting and evaluating training effectiveness.* NTSC TR-88-027. Orlando, FL: Naval Training Systems Center. (ADB129158)

Pishel, R.G., Neal, M.A., & Stapp, K.M. (1991). *Maneuver control system training effectiveness analysis.* TRAC-WSMR-TEA-91-005. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB158616)

Pleban, R.J., Brown, J.B., & Martin, M.G. (1997). *Preliminary evaluation of the computer-based tactics certification course--principles of war module.* RR 1714. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337673)

Povenmire, H.K. & Roscoe, S.N. (1972). Incremental transfer effectiveness of a ground based aviation trainer. *Human Factors,15*, 534-542.

Pugh, H.L., Parchman, S.W., & Simpson, H. (1991). *Field survey of videoteletraining systems in public education, industry, and the military.* NPRDC TR-91-7. San Diego, CA: Navy Personnel Research and Development Center. (ADA234875)

Quester, A. & Marcus, A.J. (1984). *An evaluation of the effectiveness of classroom and on-the-job training.* CNA-PP-422. Alexandria, VA: Center for Naval Analyses. (ADA112792)

Rakolta, M. J. (1994). *Network analysis of the multi-service distributed training testbed (MDT2) wide area network. Revision 1.0.* ADST/WDL/TR--94-W003419. Orlando, FL: Loral Systems ADST Program Office. (ADB198826)

Redfield, D.L. & Rousseau, E.W. (1981). A meta-analysis of experimental research of teacher questioning behavior. *Review of Educational Research, 51*, 237-245.

Resource Consultants, Inc. (1992). *Measures of crew/team training effectiveness for cost/training effectiveness analysis: Training effectiveness catalog system (TECATS) database.* Orlando, FL: Author.

Roscoe, S.N. (1971). Incremental Transfer Effectiveness. *Human Factors, 13,* 561-567.

Roscoe, S.N. (1972). A little more on incremental transfer effectiveness. *Human Factors, 14,* 363-364.

Rosenblum, D.E. (1979). *Combat effective training management study (CETRM).* Washington, DC: CETRM. (ADA101993)

Rozen, U. (1985). *Analyzing the cost effectiveness of using flight simulators in the Israeli air force.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA164864)

Russell, T.L. (1998). *The "No Significant Difference" phenomenon as reported in 248 research reports, summaries, and papers. Fourth edition.* Published on world wide web at: http://teleeducation.nb.ca/phenom. Raleigh, NC: North Carolina State University.

Salter, M.S. (1998). *Full crew interactive simulation trainer -- Bradley (FIST-B): Limited user assessment.* RR 1724. Ft. Benning, GA: U.S. Army Research Institute Field Unit. (ADA345818)

Sassone, P.G. & Schaffer, W.A. (1978). *Cost-benefit analysis: A handbook.* New York, NY: Academic Press.

Schwab, J. & Gound, D. (1988). *Concept evaluation program of simulation networking (SIMNET).* Ft. Knox, KY: U.S. Army Armor and Engineering Board. (ADB120711)

Scott, B.B., Djang, P., Laferriere, R. (1995). *Reserve component (RC) mobile close combat tactical trainer (M-CCTT) integration and deployment study.* TRAC-WSMR-TR-95-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Center. (ADB202913)

Semple, C.A. (1974). *Guidelines for implementing training effectiveness evaluations.* NTEC TR NAVTRAEQUIPCEN 72-C-0209-3. Orlando, FL: Naval Training Equipment Center.

Sheppe, M.L., Sheppard, D.J., & McDonald, R.G. (1990). *Waterfront trainer program evaluation of phase II.* TR 90-017. Orlando, FL: Naval Training Systems Center. (ADB151857)

Shlechter, T.M. Bessemer, D.W., & Kolosh, K.P. (1991). *The effects of SIMNET role-playing on the training of prospective platoon leaders.* TR 938. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA244913)

Shlechter, T.M., Bessemer, D.W., Nesselroade, P., & Anthony, J. (1995). *An initial evaluation of a simulation-based training program for Army National Guard units.* ARI RR 1679. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA297271)

Shlechter, T.M., Kraemer, R.E., Bessemer, D.W., Burnside, B.L., & Anthony, J. (1996). *Perspectives on the virtual training program from members of its initial observer/controller team.* ARI-RR-1691. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA310080)

Shute, V. J. & Gawlick-Grendell, L.A. (1992). *If practice makes perfect, what does less practice make?* AL-TP-1992-0017. Brooks AFB, TX: Armstrong Laboratory. (ADA251769)

Shute, V.J. & Psotka, J. (1994). *Intelligent tutoring systems: Past, present, and future.* AL/HR-TP-1994-0005. Brooks AFB, TX: Armstrong Laboratory. (ADA280011)

Shute, V.J. & Regian, J.W. (1993). Principles for evaluating intelligent tutoring systems. *Journal of Artificial Intelligence in Education, 4*(2/3), 245-272.

Shute, V.J. (1991). *Meta-evaluation of four intelligent tutoring systems: Promises and products.* AL-TP-1991-0040. Brooks AFB, TX: Armstrong Laboratory. (ADA243267)

Simpson, H. (1995). *Cost-effectiveness analysis of training in the department of defense.* DMDC TR 95-004. Monterey, CA: Defense Manpower Data Center. (ADA302985)

Simpson, H., West, W.D., & Gleisner, D. (1995). *The use of simulation in military training: Value, investment, and potential.* DMDC TR 95-007. Monterey, CA: Defense Manpower Data Center. (ADA303348)

Simpson, H., Wetzel, C. D., & Pugh, H. L. (1995). *Delivery of division officer navy leadership training by videoteletraining: initial concept test and evaluation.* NPRDC-TR-95-7. Navy Personnel Research and Development Center. (ADA298102)

Simutis, Z.M., Ward, J.S., Harman, J., Farr, B.J., & Kern, R.P. (1988). *ARI research in basic skills education.* RR 1486. Alexandria, VA : U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA201402)

Skog, D., Neal, M.A., & Fields, J.E. (1994). *Breacher cost and training effectiveness analysis (CTEA): Volume I - Main report.* TRAC-WSMR-CTEA-93-019-1. Volume II - Appendices. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB191580) (ADB191753)

Slavin, R.E. (1980). Cooperative learning. *Review of Educational Research, 50,* 315-342.

Smith, B.W., & Cross, K.D. (1992). *Assessment of Army Aviators' Ability to Perform Individual and Collective Tasks in the Aviation Networked Simulator (AIRNET).* RN-92-32. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA250293)

Smith, J.W. (1989). *The operational independent evaluation plan (IEP) for the close combat tactical trainer (CCTT) force development testing and experimentation (FDTE).* Ft. Leavenworth, KS: TRADOC Independent Evaluation Directorate. (ADB129761)

Smith, M.D. & Hagman, J.D. (1993). *Interdevice transfer of training between the guard unit armory device, full-crew interactive simulation trainer-armor and the mobile conduct-of-fire trainer.* RR 1635. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA263370)

Smith, S.E. & Graham, S.E. (1990). *Comparability of an armor field and simulation networking (SIMNET) performance test.* TR 895. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226353)

Solick, R.E. & Lussier, J.W. (1988). *Design of battle simulations for command and staff training.* TR 788. Ft. Leavenworth, KS: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA196655)

Solomon, H. (1986). *Economic issues in cost-effectiveness analyses of military skill training.* IDA Paper P-1897. Alexandria, VA: Institute for Defense Analyses. (AD-A171106)

Spector, P.E. (1992). *Summated rating scale construction.* Beverly Hills, CA: Sage Publications.

Sterling, B. (1996). *Relationships between platoon gunnery training and live-fire performance.* RR 1701. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA319342)

Stewart, J.E. (1994). *Using the backward transfer paradigm to validate the AH-64 simulator training research advanced testbed for aviation.* RR 1666. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285758)

Stoloff, P.H. (1991). *Cost-effectiveness of U.S. Navy video teletraining system alternatives.* CRM Research Memorandum 91-165. Alexandria, VA: Center for Naval Analyses. (ADB171320)

Taylor, H.L., Orlansky, J. Levine, D.B., Honig, J.G., & Moses, F.L. (1996). Evaluation of the performance and cost-effectiveness of the multi-service distributed training testbed (MDT2). Royal Aeronautical Society. *Conference Proceedings: The Progress and Direction of Distributed Interactive Simulation*, November 6-7, 1996.

TEXCOM (1990). *Close Combat Tactical Trainer (CCTT). Force Development Testing and Experimentation (FDTE).* TCATC-FD-0200. Ft. Hood, TX: Author. (ADB147145)

TEXCOM (1997). *Test data report for the Close Combat Tactical Trainer limited user test.* TDR-97-LUT-1645A. Ft. Hood, TX: Author. (ADB228904)

TEXCOM (1998). *Initial operational test and evaluation: Close Combat Tactical Trainer (CCTT) event design plan.* Ft. Hood, TX: Author. (ADB233901)

Thomas, B.W. & Gainer, C.A. (1990, May). Simulation networking: Low fidelity simulation in U.S. Army aviation. *Proceedings of the Royal Aeronautical Society* (pp. 18.1-18.11). London, England.

Thomas, G.S., Houck, M.R., & Bell, H.H. (1990). *Training evaluation of air combat simulation.* AL-TR-1990-3. Williams AFB, AZ: Air Force Armstrong Laboratory. (ADB145631)

Thorpe, J.A., Varney, N.C., McFadden, R.W., LeMaster, W.D., & Short, L.H. (1978). *Training effectiveness of three types of visual systems for KC-135 flight simulators.* AFHRL-TR-78-16. Brooks AFB, TX: Air Force Human Resources Laboratory. (ADA060253)

Turnage, J.J., Houser, T.L., & Hofmann, D.A. (1990). *Assessment of performance measurement methodologies for collective military training.* RN 90-126. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA227971)

Universal Cities Studios, Inc. (1985). *Back to the future.* Motion Picture. Hollywood, CA: Author.

Watson, B.L. (1992). *SIMNET-D/JANUS(T) comparison study.* TRAC-WSMR-TM-92-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Command. (ADB164784)

Wetzel, C. D., Simpson, H., & Seymour, G.E. (1995). *The use of videoteletraining to deliver chief and leading petty officer navy leadership training: evaluation and summary.* NPRDC-TR-95-8. San Diego, CA: Navy Personnel Research and Development Center. (ADA298374)

Wheaton, G. R., Rose, A. M., Fingerman, F.W., Leonard, R. L, & Boycan, G. G. (1976). *Evaluation of three burst-on-target trainers.* RM 76-18. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA076820)

Whitten, T., Horey, J., & Jones, S. (1989). *Training effectiveness evaluation of the acoustic operator trainer for the AN/SQQ-89 (V) surface antisubmarine warfare combat system, device 14E35.* TR 89-030. Orlando, FL: Naval Training Systems Center. (ADB140898)

Willett, J.B., Yamashita, J.J., & Anderson, R.D. (1983). A meta-analysis of instructional systems applied in science teaching. *Journal of Research in Science Teaching, 20,* 405-417.

Witmer, B.G. (1988). *Device-based gunnery training and transfer between the videodisk gunnery simulator (VIGS) and the unit conduct of fire trainer (U-COFT).* TR 794. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA197769)

Wood, M.A. (1987). *Final report on-site user test (OSUT) of Brigade/Battalion Battle Simulation (BBS).* Report No. USACSTA-6602. Aberdeen Proving Grounds, MD: U.S. Army Combat Systems Test Activity. (ADB117073)

Worley, R.D., Simpson, H.K., Moses, F.L., Aylward, M., Bailey, M., & Fish, D. (1996). *Utility of modeling and simulation in the department of defense: Initial data collection.* IDA Document D-1825. Alexandria, VA: Institute for Defense Analyses. (ADA312153)

APPENDIX A. REFERENCE LISTS FOR CHAPTER 3

Overview

This appendix contains 13 lists of references keyed to Chapter 3. The references on these lists provide examples of the evaluation methods described in that chapter. Some evaluations use more than one evaluation method (for example, experiment and judgment) but are placed in lists based on the primary method used in the evaluation..

Contents of the reference lists are as follows:

- A-1. True Experiment
- A-2. Pre-Experiment
- A-3. Quasi-Experiment
- A-4 . Test
- A-5. Transfer Experiment
- A-6 . Ex Post Facto
- A-7. Judgment (Users)
- A-8 . Judgment (SMEs)
- A-9 . Judgment (Analysts)
- A-10. Analysis (Evaluate)
- A-11. Analysis (Compare)
- A-12. Analysis (Optimize)
- A-13. Survey

Reference List A-1. True Experiment

Anderson, J.R. & Reiser, B.J. (1985, April). The LISP tutor. Peterborough, NH: *Byte.*

Anderson, J.R., Boyle, F.C., & Reiser, B.J. (1985). Intelligent tutoring systems. *Science, 228,* 456-462.

Banks, J.H., Hardy, G.D., Scott, T.D., Kress, G., & Word, L.E. (1977). *REALTRAIN validation for rifle squads: Mission accomplishment.* RR 1192. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA043515)

Bierbaum, C.R., McAnulty, D. M., Cross, K.D. (1992). *Effectiveness of contractor mission instructors in the 160th special operations aviation regiment basic mission qualification course.* RN-92-30. Ft. Rucker, AL: Anacapa Sciences, Inc. (ADA251845)

Brown, R.E., Pishel, R.G., & Southard L.D. (1988). *Simulation networking (SIMNET) preliminary training developments study.* TRAC-WSMR-TEA-8-99. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB120874)

Butler, W.G. (1982). *Training developments study - Bradley fighting vehicle unit conduct of fire trainer.* TRASANA TEA 28-82. Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADB954498)

Butler, W.G., Reynolds, M.J., Kroh, M.Z., & Thorne, H.W. (1982). *Training developments study--M1 (Abrams) tank unit conduct of fire trainer.* TRASANA TEA 11-82. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB954521)

Campshure, D.A., Witmer, B.G., & Drucker, E.H. (1990). *The effects of amount of M1 unit conduct of fire trainer (U-COFT) transition training on crew gunnery proficiency.* RN 90-03. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA219924)

Dannhaus, D. (1980). *Multiple launch rocket system (MLRS) cost and training effectiveness analysis (CTEA).* TRASANA Report No. TEA-380. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB077491)

Dannhaus, D. (1980). *Multiple launch rocket system (MLRS) cost and training effectiveness analysis (CTEA) preliminary (phase I).* TRASANA CTEA-3-80. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB045320)

Dannhaus, D.M., Hughes, C.R., & Shea, J.M. (1982). *Multiple integrated laser engagement system (MILES) air ground engagement simulation/air defense (AGES/AD) cost and training effectiveness analysis (CTEA).* TRASANA TEA 12-82. White Sands Missile Range, NM: U.S. Army Systems Analysis Activity. (ADB954527)

Dyer, J.L., Westergren, A.J., Shorter, G.W., & Brown, L.D. (1997). *Combat vehicle training with thermal imagery.* TR 1074. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA342559)

Eisley, M.E., Hagman, J.D., Ashworth, R.L., & Viner., M.P. (1990). *Training effectiveness evaluation of the squad engagement training system (SETS).* RR 1562. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226406)

Fletcher, J.D., Hawley, D.E., & Piele, P.K. (1990). Costs, effects, and utility of microcomputer assisted instruction in the classroom. *AERA Journal, 27*(4), 783-806.

Gardner, J. (1988). *A comparison of the national training center and the JANUS (T) combat model battle results.* Monterey, CA: Naval Postgraduate School. (ADA200115)

Graham, S.E., Shlechter, T.M., & Goldberg, S.L. (1986). *A preliminary evaluation of a model maintenance training program for reserve component units.* RR 1421: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA173909)

Greene, D.E. & Haynes, M.D. (1988). *Concept evaluation program (CEP) test of the tube-launched, optically tracked, wire-guided (TOW) training strategy.* USAIB Project No. 3901. Ft. Benning, GA: U.S. Army Infantry Center. (ADB125492)

Greene. D.E. & Tolbert, R.J. (1988). *Customer test of TOW precision gunnery training system (PGTS).* USAIB Project Number 3869. Ft. Benning, GA: U.S. Army Infantry Board. (ADB129284)

Hamel, C.J., Kincaid, J.P., & Thompson, J. (1982). *Field test of a numerical basic skills curriculum.* TR 135. Orlando, FL: Training Analysis and Evaluation Group. (ADA122354)

Hamilton, D.B. (1991). *Training effectiveness of the AH-64A combat mission simulator for sustaining gunnery skills.* RR 1604. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA244820)

Hendrix, C.L. (1995). *A comparison of one-way video and two-way video educational videoteleconferencing.* Thesis. Wright-Patterson AFB, OH: Air Force Institute of Technology. (ADA294654)

Hoffman, R.G. & Melching, W.H. (1984). *Field trials of the MK60 tank gunnery simulator in armor institutional training courses, Volume 1: Final report.* RR 1381. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA170939)

Howard, F.H., Henry, J.B., Kinney, P., & Dannhaus, D. (1991). *Distributed training strategy training effectiveness analysis MOS 63W desktop video pilot study.* TRAC-WSMR-TEA-91-023. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB163581)

Hughes, C., Butler, W., Sterling, B., & Berglund, A. (1987). *M1 unit conduct-of-fire trainer (UCOFT) post fielding training effectiveness analysis (PFTEA).* White Sands Missile Range, NM: TRADOC Analysis Command. (ADB113298)

Johnson, D.M. & Wightman, D.C. (1995). *Using virtual environments for terrain familiarization: Validation.* RR 1686. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA304416)

Kincaid, J.P., Swope, W.M., Brown, C.J., Pereyra, B., & Thompson, J. (1982). *Field test of the verbal skills curriculum.* TR 128. Orlando, FL: Training Analysis and Evaluation Group. (ADA118875)

King, F.J. (1982). *Evaluation of a videodisc delivery system for teaching students to troubleshoot the AN/VRC-12 medium-powered radio series.* Report TDI-TR-82-7. Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA157890)

Landers, M.D., Hunt, K.T. (1991). *Guard unit armory device full-crew interactive simulation trainer for armor (GUARD FIST I) customer test.* 91-CT-990. Ft. Knox, KY: TEXCOM Armor and Engineer Board. (ADB154266)

Lassiter, D.L., Vaughn, J.S., Smaltz, V.E., & Morgan, B.B. (1990). A comparison of two types of training interventions on team communication performance. *Paper presented at 1990 Meeting of the Human Factors Society*, Orlando, FL.

Lesgold, A. (1994). Assessment of intelligent training technology. In E.L. Baker & H.F. O'Neil (Eds.). *Technology Assessment in Education and Training* (pp. 97-116). Hillsdale, NJ: Lawrence Erlbaum.

Lickteig, C.W. & Burnside, B.L. (1986). *Land navigation skills training: an evaluation of computer and videodisc-based courseware.* ARI-TR-729. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA181031)

Lickteig, C.W. & Burnside, B.L. (1987). *Remedial skills training: An evaluation of computer and videodisc-based courseware.* ARI-RR-1444. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA190832)

Lickteig, C.W. & Collins, J.W. (1995). *Combat vehicle command and control system evaluation: Vertical integration of an armor battalion.* TR 1021. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA292718)

Malec, V.M. & Lusczak, M. (1987). *Field evaluation of interactive simulation for maintenance training: SH-3 helicopter electro-mechanical bladefold system.* NPRDC TR-88-2. San Diego, CA: Navy Personnel Research and Development Center. (ADA185923)

Martellaro, H.C., Thorne, H.W., Bryant, J.A., & Pierce, M.A. (1985). *Tank gunnery/conduct of fire trainer (COFT) M1 training effectiveness analysis (TEA).* TRASANA-TEA-23-85. White Sands Missile Range, NM: U.S. Army Training and Doctrine Command Systems Analysis Activity. (ADB097355)

McAnulty, D.M. (1992). *Effectiveness of the AH-1 flight and weapons simulator for sustaining aerial gunnery skills.* ASI690-343-91. Ft. Rucker, AL: Anacapa Sciences, Inc. (ADA250810)

McDonald, B., Hurlock, R., Ellis, J., & Whitehill, B. (1989). *Self-paced and group-paced instruction in basic electronics and electricity training.* NPRDC TR 89-8. San Diego, CA: Navy Personnel Research and Development Center. (ADB132642)

Merrill, D.C., Reiser, B.J., & Merrill, S.K. (1995). *Tutoring: Guided learning by doing.* RAND/RP-329. Santa Monica, CA: Rand. (ADA304686)

Millard, S.L. (1986). An assessment of the effectiveness of training helicopter initial entry students in simulators. In K.D. Cross & S.M. Szabo (Eds.), *Human factors research in aircrew performance and training: Final summary report.* RN 86-94, pp. 105-108. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA175348)

Moon, J. & Strassel, H. (1982). *Independent evaluation report (IER) for M2/M3 unit conduct of fire trainer (U-COFT).* USAIS Report No. IER 66519. Ft. Benning, GA: Directorate of Training Developments, U.S. Army Infantry School. (ADB068136)

Palmer, R.L. (1990). *Single-channel ground and airborne radio system (SINCGARS) operator training evaluation.* RR 1579. Ft. Hood, TX: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA232522)

Pfeiffer, M.G. (1990). *Training effectiveness evaluation of interactive videodisc courseware (device 21H15).* TR 90-011. Orlando, FL: Naval Training Systems Center. (ADB149460)

Phelps, R.H., Wells, R.A., Ashworth, R.L., & Hahn, H.A. (1991). Effectiveness and costs of distance education using computer-mediated communication. *American Journal of Distance Education, 5*(3), 7-19.

Powers, T.R., McCluskey, M.R., Haggard, D.F., Boycan, G.G., & Steinheiser, F. (1974). *Determination of the contribution of live firing to weapons proficiency.* Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA036060)

Rapkoch, J.M. & Robinson, F.D. (1986). *Concept evaluation program of gunnery training devices.* Final Report 6CEP342. Ft. Knox, KY: U.S. Army Armor and Engineer Board. (ADB104075)

Root, T.R., Hayes, J.F., Word, L.E., Shriver, E.L., & Griffin, G.R. (1979). *Field test of techniques for tactical training of junior leaders in infantry units (Project EFFTRAIN).* TR 79-A21. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA075604)

Rupinski, T.E. & Stoloff, P.H. (1990). *An evaluation of navy video teletraining (VTT).* CRM 90-36. Alexandria, VA: Center for Naval Analyses. (ADA239180)

Rupinski, T.E. (1991). *Analysis of video teletraining utilization, effectiveness, and acceptance.* CRM Research Memorandum 91-159. Alexandria, VA: Center for Naval Analyses. (ADB171314)

Scholtes, T.G. & Stapp, K.M. (1994). *Engagement skills trainer (EST) training effectiveness analysis (TEA).* TRAC-WSMR-TEA-94-018. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB196582)

Schwab, J.R. & Gound D. (1988). *Concept evaluation of simulation networking (SIMNET).* TR 86-CEP345. Ft. Knox, KY: U.S. Army Armor and Engineer Board. (ADB120711)

Shlechter, T.M. (1988). *The effects of small group and individual computer-based instruction on retention and on training lower ability soldiers.* RR 1497. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA203793)

Shute, V. J. & Gawlick-Grendell, L. A. (1994). What does the computer contribute to learning? *Computers and Education: An International Journal, 23*(3), 177-186.

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments(1)*, 51-76.

Simpson, H., Pugh, H. L., Parchman, S. W. (1992). *The use of videoteletraining to deliver hands-on training: concept test and evaluation.* NPRDC-TN-92-14. San Diego, CA: Navy Personnel Research and Development Center. (ADA250708)

Simpson, H., Pugh, H.L., & Parchman, S.W. (1991). *Empirical comparison of alternative video teletraining technologies.* TR-92-3. San Diego, CA: Navy Personnel Research and Development Center. (ADA242200)

Simpson, H., Wetzel, C. D., & Pugh, H. L. (1995). *Delivery of division officer navy leadership training by videoteletraining: initial concept test and evaluation.* NPRDC-TR-95-7. Navy Personnel Research and Development Center. (ADA298102)

Singer, M.J., Allen, R.C., McDonald, D.P., & Gildea, J.P. (1997). *Terrain appreciation in virtual environments: Spatial knowledge acquisition.* TR 1056. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA325520)

Smith, M.D. (1998). *Assessment of the SIMITAR gunnery training strategy through development of a database of gunnery outcome measures.* RR 1721. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA344930)

Smith, S.E. & Graham, S.E. (1990). *Comparability of an armor field and simulation networking (SIMNET) performance test.* TR 895. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226353)

Stout, R.J., Salas, E., & Fowlkes, J.E. (1997). Enhancing teamwork in complex environments through team training. *Group Dynamics: Theory, Research, and Practice, 1*(2), 169-182.

U.S. Army Infantry Board (1985). *Concept evaluation program (CEP) test of the Bradley infantry fighting vehicle gowen south.* TRADOC Project No. 4-CEP 179. Ft. Benning, GA: Author. (ADB091572)

Watson, B.L. (1992). *SIMNET-D/JANUS(T) comparison study.* TRAC-WSMR-TM-92-009. White Sands Missile Range, NM: U.S. Army TRADOC Analysis Command. (ADB164784)

Wetzel, C. D., Pugh, H. L., Van Matre, Nick, Parchman, Steven W. (1996). *Videoteletraining delivery of a quality assurance course with a computer laboratory.* NPRDC-TR-96-6. San Diego, CA: Navy Personnel Research and Development Center. (ADA308013)

Wetzel, C. D., Simpson, H., & Seymour, G.E. (1995). *The use of videoteletraining to deliver chief and leading petty officer navy leadership training: evaluation and summary.* NPRDC-TR-95-8. San Diego, CA: Navy Personnel Research and Development Center (ADA298374)

Wetzel, C.D. (1995). *Evaluation of a celestial navigation refresher course delivered by videoteletraining.* NPRDC-TR-96-2. San Diego, CA: Navy Personnel Research and Development Center. (ADA300925)

Wetzel, C.D., Radtke, P.W., Parchman, S.W., & Seymour, G.W. (1996). *Delivery of a fiber optic cable repair course by videoteletraining.* TR-96-4. San Diego, CA: Navy Personnel Research and Development Center. (ADA304318)

Wetzel-Smith, S.K., Ellis, J.A., Reynolds, A.M., & Wulfeck, W.H. (1995). *The interactive multisensor analysis training (IMAT) system: An evaluation in operator and tactician training.* NPRDC TN-96-3. San Diego, CA: Navy Personnel Research and Development Center. (ADA302908)

Wilhoite, B.K. (1993, December). Bytes vs. Bullets: Crew-served weapons simulation based training. *Proceedings of the 15th. Interservice/Industry Training Systems and Education Conference* (pp. 475-480). Orlando, FL.

Wilkinson, G.L. (1983). *An evaluation of equipment-independent maintenance training by means of a microprocessor-controlled videodisc delivery system.* TRI-TR-83-1. Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA127011)

Wilkinson, G.L. (1985). *Evaluation of the effectiveness and potential application of the interactive videodisc instructional delivery system within the training of SIGINT/EW systems repairers, CMF33.* Ft. Eustis, VA: U.S. Army Communicative Technology Office and Fort Devens, MA: U.S. Army Intelligence School. (ADA157942)

Winkler, J.D. & Polich, J.M. (1990). *Effectiveness of interactive videodisc in Army communications training.* Santa Monica, CA: RAND Corporation. (ADA236867)

Wisher, R.A., Priest, A.N., & Glover, E.C. (1997). *Audio teletraining for unit clerks: A cost-effectiveness analysis.* RR 1712. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337689)

Reference List A-2. Pre-Experiment

Anderson, J. R. (1990). Analysis of student performance with the LISP tutor. In N. Fredericksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.). *Diagnostic Monitoring of Skill and Knowledge Acquisition.* Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. *Cognitive Science, 13*(4), 467-505.

Brown, R.E., Mullis, C.W., & Coffey, B.A. (1987). *Brigade/battalion simulation onsite user test.* TRAC-WSMR-TEA-31-87. White Sands Missile Range, NM: U.S. Army Training and Doctrine Command. (ADB117569)

Cordell, C.C., Nutter, R.V., & McDaniel, W.C. (1983). *Training effectiveness evaluation (TEE) of the advanced fire fighting training system.* TR 142. Orlando, FL: Training Analysis and Evaluation Group. (ADA126193)

Graham, S.E. (1987) *Field evaluation of a computer-based maintenance training program for reserve component units.* RR 1461: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA193085)

Hahn, C.P., Krug, R.E., & Stoddart, S.C. (1985). *Evaluation of the McFann, Gray & Associates BSEP II curriculum.* RN 85-86. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA160508)

Hahn, C.P., Krug, R.E., Rosenbaum, H., Stoddart, S.C., & Harmon, J. (1986). *Evaluation of the U.S. Army basic skills education program.* TR 277. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA178650)

Hanley, M.L. (1985). *Results of the part-task shiphandling trainer pre-prototype training effectiveness evaluation (TEE).* NAVTRAEQUIPCEN 83-C-0015-1. Orlando, FL: Naval Training Equipment Center. (ADA154409)

Harman, J., Bell, S.A., & Laughy, N. (1989). *Evaluation of the hand-held mathematics tutor.* RR 1509. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA207157)

Harris, K. (1995). *Battle force tactical training (BFTT) developmental test IIA (DT-IIA) operational test report.* Unpublished test report. Port Hueneme, CA: Naval Surface Warfare Center.

Harris, K. (1996). *Battle force tactical training (BFTT) developmental test IIB (DT-IIB) developmental test report.* Unpublished test report. Port Hueneme, CA: Naval Surface Warfare Center.

Horey, J.D. & Dwyer, D.J. (1992). *Training effectiveness evaluation of the part task trainer for the A-6E SLAM weapon system.* TR 92-014. Orlando, FL: Naval Training Systems Center. (ADB168986)

Jentsch, K.S. (1996). *BFTT training effectiveness: Lessons learned.* Briefing slides. Orlando, FL: Naval Air Warfare Center Training Systems Division.

Kraemer, R.E. & Smith, S.E. (1990). *Soldier performance using a part-task gunnery device (TOPGUN) and its effects on institutional-conduct of fire (I-COFT) proficiency.* RR 1570. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA227403)

Lampton, D.R. (1989). *Evaluation of a low fidelity battle simulation for training and evaluating command, control, and communications (C3I) skills for the armor platoon leader.* RR 1521. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA210606)

Lesgold, A., Eggan, G., Katz, S., & Rao, G. (1992). *Possibilities for assessment using computer-based apprenticeship environments.* In J.W. Regian & V.J. Shute (Eds.). Cognitive Approaches to Automated Instruction (pp. 49-80). Hillsdale, NJ: Lawrence Erlbaum.

Moon, J. & Strassel, H. (1982). *Independent evaluation report (IER) for M2/M3 unit conduct of fire trainer (U-COFT).* USAIS Report No. IER 66519. Ft. Benning, GA: Directorate of Training Developments, U.S. Army Infantry School. (ADB068136)

Nau, K.L. & Harris, W.B. (1995). *Berlin brigade squad engagement training system training effectiveness analysis (SETS TEA).* TRAC-WSMR-TEA-95-019. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB206341)

Noble, J. (1983). *Army training battle simulation system (ARTBASS) operational test II (OT-II) training developments study.* TRASANA-TEA-35-83. White Sands Missile Range, NM: U.S. Army Training and Doctrine Command Systems Analysis Activity. (ADB084732)

Pleban, R.J., Brown, J.B., & Martin, M.G. (1997). *Preliminary evaluation of the computer-based tactics certification course--principles of war module.* RR 1714. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337673)

Smith, M.D. (1998). *Assessment of the SIMITAR gunnery training strategy through development of a database of gunnery outcome measures.* RR 1721. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA344930)

TEXCOM (1990). *Close combat tactical trainer (CCTT) force development testing and experimentation.* TCATC Test Report No. FD 0200. Ft. Hood, TX: Author. (ADB147145)

U.S. Army TRADOC Systems Analysis Activity (1977). *Redeye weapons system training effectiveness analysis.* TR 20-77. White Sands Missile Range, NM: Author. (ADB112051)

Wick, D.T., Millard, S.L., Cross, K.D., (1986). *Evaluation of a revised individual ready reserve (IRR) aviator training program.* ASI479-058-85. Ft. Rucker, AL: Anacapa Sciences, Inc. (ADA173811)

Reference List A-3. Quasi-Experiment

Bessemer, D.W. (1991). *Transfer of SIMNET training in the armor officer basic course.* TR 920. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA233198)

Dwyer, D.J., Fowlkes, J., Oser, R.L., & Salas, E. (1996). Panel on multi-service distributed training testbed, DIS training of military teams/groups: Case study results using distributed interactive simulation for close air support. *Proceedings of the 1996 International Training Equipment Conference.* The Hague, Netherlands, 371-380.

Dwyer, D.J., Oser, R.L., Salas, E., & Fowlkes, J.E. (1997). *Performance measurement in distributed environments: Initial results and implications for training.* Manuscript.

Kaempf, G.L., Cross, K.D., & Blackwell, N.J. (1989). *Backward transfer and skill acquisition in the AH-1 flight and weapons simulator.* Anacapa Sciences Report AS1690-312-88. Ft. Rucker, AL: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA213432)

Kraemer, R.E. & Wong, D.T. (1992). *Evaluation of a prototype platoon gunnery trainer (PGT) for armor officer basic course training.* RR 1620. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA254289)

Moses, F. L. (1996). Panel on multi-service distributed training testbed, DIS training of military teams/groups: The challenge of distributed training. *Proceedings of the 1996 International Training Equipment Conference.* The Hague, Netherlands, 358-364.

Orlansky, J., Taylor, H.L., Levine, D.B., & Honig, J.G. (1997). *The cost and effectiveness of the multi-service distributed training testbed (MDT2) for training close air support.* IDA Paper P-3284. Alexandria, VA: Institute for Defense Analyses. (ADA327227)

Oser, R.L., Dwyer, D.J., & Fowlkes, J. (1995). Team performance in multi-service distributed interactive simulation exercises: Initial results. *Proceedings of the 17th. I/ITSEC Conference.* Albuquerque, NM, 163-171.

Pfeiffer, M.G. & Rankin, W.C. (1986). *Training effectiveness evaluation of passive acoustic analysis trainer (Device 21H14).* TR 86-019. Orlando, FL: Naval Training Systems Center. (ADB108292)

Shlechter, T.M., Bessemer, D.W., Nesselroade, P., & Anthony, J. (1995). *An initial evaluation of a simulation-based training program for Army National Guard units.* RR 1679. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA297271)

Taylor, H.L., Orlansky, J. Levine, D.B., Honig, J.M., & Moses, F.C. (1996). Evaluation of the performance and cost-effectiveness of the multi-service distributed training testbed (MDT2). Royal Aeronautical Society. *Conference Proceedings: The Progress and Direction of Distributed Interactive Simulation,* November 6-7, 1996.

Whitten, T., Horey, J., & Jones, S. (1989). *Training effectiveness evaluation of the acoustic operator trainer for the AN/SQQ-89 (V) surface antisubmarine warfare combat system, device 14E35.* TR 89-030. Orlando, FL: Naval Training Systems Center. (ADB140898)

Reference List A-4. Test

Acchione-Noel, S.C., Pierce, M.A., Keaton, M.L., Mullis, C.W., Brown, R.E., Coffey, B.A., & Stapp, M. (1987). *Wheel vehicle maintenance training effectiveness analysis: Phase I commercial utility cargo vehicle.* TRAC-WSMR-TEA-26-87. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB117451)

Cassady, P.D., Laverne, J.E., Kestner, R.R., Risser, R.C., & Sasaki, R.T. (1982). *Multiple launch rocket system (MLRS) cost and training effectiveness analysis.* TRASANA TEA-56-82. Ft. Monroe, VA: U.S. Army TRADOC Systems Analysis Activity. (ADB079368)

Clark, D.R. (1989). *Independent assessment report for the Dragon precision gunnery training system (PGTS).* AMSTE-TA-W. Aberdeen Proving Grounds, MD: Headquarters, U.S. Army Test and Evaluation Command. (ADB138024)

Ennis, J.J. & Gardner, C.V. (1990). *Chaparral/FLIR post fielding training effectiveness analysis (PFTEA).* FB TEA 1-89. Ft.Bliss, TX: U.S. Army Air Defense Artillery School. (ADB156909)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990). *Verification and validation of SIMNET-T.* K/DSRD-117. Oak Ridge, TN: Martin Marietta Energy Systems, Inc. (ADB147355)

Hartley, D.S., Quillinan, J.D., & Kruse, K.L. (1990). *Verification and Validation of SIMNET-T. Phase 1.* K/DSRD-116. Oak Ridge, TN: Martin Marietta Energy Systems, Inc. (ADB147354)

Hires, J. (1990). *Independent assessment report of the guard unit armory device for full-crew interactive simulation training for the artillery (GUARD FIST II) (type classification low rate production).* Aberdeen Proving Ground, MD: U.S. Army Test and Evaluation Command. (ADB151552)

Pate, D., Lewis, B.D., & Wolf, G. (1988). *Innovative test of the simulator network (SIMNET) system.* Ft. Bliss, TX: U.S. Army Air Defense Artillery Board. (ADB124036)

Pishel, R.G., Neal, M.A., & Stapp, K.M. (1991). *Maneuver control system training effectiveness analysis.* TRAC-WSMR-TEA-91-005. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB158616)

Salter, M.S. (1998). *Full crew interactive simulation trainer -- Bradley (FIST-B): Limited user assessment.* RR 1724. Ft. Benning, GA: U.S. Army Research Institute Field Unit. (ADA345818)

Smith, B. & Cross, K. (1992). *Assessment of Army aviator's ability to perform individual and collective tasks in the aviation networked simulator (AIRNET).* RN 92-32. Ft. Rucker, AL: Anacapa Sciences, Inc. (ADA250293)

Sterling, B.S. & Brett, D.C. (1990). *AT4 post fielding training effectiveness analysis.* TRAC-WSMR-TEA-90-034. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB158918)

Sterling, B.S. & Hansen, A.D. (1990). *OH-58D observation helicopter post fielding training effectiveness analysis.* TRAC-WSMR-TEA-90-022. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB161702)

TEXCOM (1997). *Test data report for the Close Combat Tactical Trainer limited user test.* TDR-97-LUT-1645A. Ft. Hood, TX: Author. (ADB228904)

Vertical Flight Aircraft Joint Venture (1996). *Summary evaluation report USMC H-1 RP3S aviation survival refresher training.* Orlando, FL: Naval Air Warfare Center Training Systems Division. Author.

Wood, M.A. (1987). *Final report on-site user test (OSUT) of Brigade/Battalion Battle Simulation (BBS).* Report No. USACSTA-6602. Aberdeen Proving Grounds, MD: U.S. Army Combat Systems Test Activity. (ADB117073)

Reference List A-5. Transfer Experiment

Bauer, R.W. (1978). *Training transfer from mini-tank range to tank main gun firing.* Technical Paper 285. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA061566)

Browning, R.F., McDaniel, W.C., Scott, P.G., & Smode, A.F. (1982). *An assessment of the training effectiveness of device 2F64C for training helicopter replacement pilots.* TR 127. Orlando, FL: Training Analysis and Evaluation Group. (ADA118942)

Hagman, J.D. & Smith, M.D. (1991). Device-based prediction of tank gunnery performance. *Military Psychology, 8*(2), 59-68.

Hart, R.J., Hagman, J.D., & Bowne, D.S. (1990). *Tank gunnery: Transfer of training from TOPGUN to the conduct-of-fire trainer.* RR 1560. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA223165)

Hughes, C., Morales-Steigely, M., & Musser, M. (1990). *M2/M3 unit conduct-of-fire trainer (UCOFT) post fielding training effectiveness analysis (PFTEA).* White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157015)

Kaempf, G.L. & Blackwell, N.J. (1990). *Transfer of training study of emergency touchdown maneuvers in the AH-1 flight and weapons simulator.* RR 1561. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226360)

Kaempf, G.L. (1986). Backward transfer of emergency touchdown maneuvers. In K.D. Cross & S.M. Szabo (Eds.), *Human factors research in aircrew performance and training: Final summary report.* RN 86-94, pp. 42-51. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA175348)

Kaempf, G.L., Cross, K.D., & Blackwell, N.J. (1989). *Backward transfer and skill acquisition in the AH-1 flight and weapons simulator.* Anacapa Sciences Report AS1690-312-88. Ft. Rucker, AL: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA213432)

McDaniel, W.C. (1987). *Training effectiveness evaluation of aviation antisubmarine warfare basic operator trainer (Device 14D1).* TR 87-019. Orlando, FL: Naval Training Systems Center. (ADB118742)

Nullmeyer, R.T. & Rockway, M.R. (1993?). *Effectiveness of the C-130 weapon system trainer for tactical aircrew training.* Williams AFB, AZ: Air Force Human Resources Laboratory

Pfeiffer, M.G. & Dwyer D.J. (1991). *Training effectiveness of the F/A-18 weapon tactics trainer (Device 2E7).* TR 91-1008. Orlando, FL: Naval Training Systems Center. (ADB160186)

Povenmire, H.K. & Roscoe, S.N. (1971). An evaluation of ground-based flight trainers in routine primary flight training. *Human Factors, 13*(2), 109-116.

Rose, A.M., Wheaton, G. R., Leonard, R.L., Fingerman, F.W., & Boycan, G.G. (1976). *Evaluation of two gunnery trainers.* RM 76-19. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA082954)

Schendel, J.D., Heller, F.H., Finley, D.L., & Hawley J.K. (1984). *Use of weaponeer marksmanship trainer in predicting M16A1 rifle qualification performance.* RR 1370. Ft Benning, GA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA156805)

Shute, V. J. and Gawlick-Grendell, L.A. (1992). *If practice makes perfect, what does less practice make?* AL-TP-1992-0017. Brooks AFB, TX: Armstrong Laboratory. (ADA251769)

Shute, V. J., Regian, J. W., & Gawlick, L.A. (1996). *Modeling practice, performance, and learning.* AL-TP-1995-0039. Brooks AFB, TX: Armstrong Laboratory. (ADA306162)

Smith, M.D. & Hagman, J.D. (1993). *Interdevice transfer of training between the guard unit armory device, full-crew interactive simulation trainer-armor and the mobile conduct-of-fire trainer.* RR 1635. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA263370)

Stewart, J.E. (1994). *Using the backward transfer paradigm to validate the AH-64 simulator training research advanced testbed for aviation.* RR 1666. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA285758)

Thorpe, J.A., Varney, N.C., McFadden, R.W., LeMaster, W.D., & Short, L.H. (1978). *Training effectiveness of three types of visual systems for KC-135 flight simulators.* AFHRL-TR-78-16. Brooks AFB, TX: Air Force Human Resources Laboratory. (ADA060253)

Turnage, J.J. & Bliss J.P. (1990). *An analysis of skill transfer for tank gunnery performance using TOPGUN, VIGS, and ICOFT trainers.* TR 916. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA240628)

Wheaton, G. R., Rose, A. M., Fingerman, F.W., Leonard, R. L, & Boycan, G. G. (1976). *Evaluation of three burst-on-target trainers.* RM 76-18. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA076820)

Witmer, B.G. (1988). *Device-based gunnery training and transfer between the videodisk gunnery simulator (VIGS) and the unit conduct of fire trainer (U-COFT).* TR 794. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA197769)

Reference List A-6. Ex Post Facto

Bessemer, D.W. (1991). *Transfer of SIMNET training in the armor officer basic course.* TR 920. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA233198)

Campshure, D.A. & Drucker, E.H. (1990). *Predicting first-run gunnery performance on tank table VIII.* RR 1571. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA228201)

Derrick, D.L. & Davis, M.S. (1993). *Cost-effectiveness analysis of the C-130 aircrew training system.* AL-TR-1992-0173. Williams Air Force Base, AZ: Armstrong Laboratory. (ADB171592)

Hall, E.R. & Freda, J.S. (1982). *A comparison of individualized and conventional instruction.* TR 117. Orlando, FL: Training Analysis and Evaluation Group. (ADA115319)

Hunt, J.P., Parish, J.R., Martere, R.F., & Evans, K.L. (1987). *A cost and training effectiveness analysis (CTEA) of moving target engagement training programs for the M16A1 rifle.* RN 87-40. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA186821)

O'Brien, R.M. (1989). *A comparison of training effectiveness of formal and on-the-job enlisted rate training in the United States Coast Guard.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA223820)

Orlansky, J., Metzko, J., Morrison, J., & Pickell, G. (1997). *Assessment of SIMITAR: Status report one.* IDA Document D2069. Alexandria, VA: Institute for Defense Analyses. (ADA331447)

Pfeiffer, M.G. & Guynn, S.J.. (1986). *Training effectiveness evaluation of close-in weapon system maintenance trainer (Device 11G2).* TR 86-018. Orlando, FL: Naval Training Systems Center. (ADB108298)

Rivera, F.G. & Nantze, S.R. (1991). *Air battle captain (ABC) cost and training effectiveness analysis.* TRAC-WSMR-CTEA-91-026. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB161285)

Rivera, F.G. (1997). *University of North Dakota (UND) rotary-wing training (RWT) program cost and training effectiveness analysis (CTEA) technical report.* TRAC-WSMR-CTEA-97-004. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB219903)

Shlechter, T.M. Bessemer, D.W., & Kolosh, K.P. (1991). *The effects of SIMNET role-playing on the training of prospective platoon leaders.* TR 938. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA244913)

Spendley, J.K. (1990). *Effectiveness of the U.S. Navy's basic skills enhancement program entitled functional applied skills training (FAST).* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA241804)

Sterling, B. (1996). *Relationships between platoon gunnery training and live-fire performance.* RR 1701. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA319342)

Swope, E.H., Copeland, D.R, & Kincaid, J.P. (1982). *Cost/benefit evaluation of three English language training programs for potential Navy use.* TAEG Report 134. Orlando, FL: Navy Training Analysis and Evaluation Group. (ADA122445)

Zamarripa, A.A., Nantze, S.R., Lascelles, K., Coffey, B.A., & Catherson, N.S. (1994). University of North Dakota (UND) rotary-wing training (RWT) program cost and training effectiveness analysis (CTEA) final report. TRAC-WSMR-CTEA-95-008. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB204711)

Reference List A-7. Judgment (Users)

Barber, H.G. & Solick, R.E. (1980). *MILES training and evaluation test, USAREUR: battalion command group training.* RN 1290. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA110503)

Brown, R. & Mullis, C. (1988). *Simulation networking assessment of perceptions - I.* TRASANA Report No. LR-1-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB118627)

Brown, R. & Mullis, C. (1988). *Simulation networking assessment of perceptions - II.* TRASANA Report No. LR-2-88. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB146645)

Crane, P.M. & Berger, S.C. (1993). *Multiplayer simulator based training for air combat.* Williams AFB, AZ: Air Force Armstrong Laboratory.

Evans, R.M. & Braby, R. (1983). *Self-paced and conventional instruction in Navy training: A comparison on elements of quality.* TAEG TR-147. Orlando, FL: Training Analysis and Evaluation Group. (ADA132402)

Fletcher, J.D. (1988). *Responses of the 1/10 cavalry to SIMNET.* IDA Analysis Memorandum No. M-494. Arlington, VA: Defense Sciences Office. (ADA200449)

Hoffman, G.R. (1997). *Combat support and combat service support expansion to the virtual training program SIMNET battalion exercise: History and lessons learned.* RR 1717. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA341201)

Houck, M.R., Thomas, G.S., & Bell, H.H. (1991). *Training evaluation of the F-15 advanced air combat simulation.* AL-TP-1991-0047. Williams AFB, AZ: Air Force Armstrong Laboratory. (ADA241675)

Lyon, D.K., Mullis, H.E., Baxley, C.R., Hooper, T.L., & Bickings, D.K. (1979). *Multiple integrated laser engagement system (MILES) operational test II (Miles OT II).* OT 210. Ft. Leavenworth, KS: Combined Arms Training Developments Activity. (ADB039450)

McDonald, G.W., Broeder, R.F., & Cutak, R.J. (1989). Multiship air combat simulation. *Proceedings of the Interservice/Industry Training Systems Conference* (pp. 148-158).

Mirabella, A. (1995). Symposium on distributed simulation for military training of teams/groups: MDT2 system assessment and effectiveness. *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting,* San Diego, CA, 1321-1325.

Mirabella, A., Sticha, P., & Morrison, J. (1997). *Assessment of user reactions to the multi-service distributed training testbed (MDT2) system.* TR 1061. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA328473)

Shlechter, T.M., Shadrick, S.B., Bessemer, D.W., & Anthony, J. (1997). *An examination of training issues associated wtih the virtual training program.* TR 1072. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA338732)

Starkel, R. (1997, April 8). *East coast demonstration surveys* [Battle Force Tactical Trainer]. Memorandum. Washington, DC: BFTT Program Office.

Thomas, G.S., Houck, M.R., & Bell, H.H. (1990). *Training evaluation of air combat simulation.* AL-TR-1990-3. Williams AFB, AZ: Air Force Armstrong Laboratory. (ADB145631)

Reference List A-8. Judgment (SMEs)

Bryant, J.A., Lewis, N.L., Stapp, M., Zamarripa, A.A., Cox, J., Wilhelm, J., & Walker, M. (1992). *JANUS(A) brigade/battalion simulation cost and training effectiveness analysis.* TRAC-WSMR-CTEA-92-006. TRADOC Analysis Command, White Sands Missile Range, NM. (ADB186373)

Clark, D. (1991). *Independent assessment report for the TOW gunner trainer precision gunnery training system.* Aberdeen Proving Grounds, MD: Headquarters, U.S. Army Test and Evaluation Command. (ADB160229)

Holstead, J. (1989). *Large scale simulation networking (SIMNET) operational effectiveness appraisal (OEA).* TAC Project 89-190T. Eglin Air Force Base, FL: USAFTWAC. (ADB133378)

Keller, A.R., Parrish, J.J., Harrison, J.A., & Macklin, L. (1992). *Mobile aircrew sustainment training-Apache (MAST-A) training effectiveness analysis - Phase I (TEA-I).* Report No. AVNC-DOS-92-5. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB172118)

Kelly, J.F. (1995). *A training effectiveness evaluation of leathernet.* Thesis. Monterey, CA: U.S. Naval Postgraduate School. (ADA303556)

Kendig, L.D. & McCollum, S.A. (1992). *UH-1 helicopter training effectiveness assessment.* Ft. Hood, TX: U.S. Army Test and Experimentation Command. (ADB169327)

Mullis, C.W. (1991). *Brigade/battalion simulation preliminary training developments study – Update.* TRAC-WSMR-TEA-91-024. White Sands Missile Range, NM: U.S. Army Training and Doctrine Command. (ADB158630)

Quester, A. & Marcus, A.J. (1984). *An evaluation of the effectiveness of classroom and on-the-job training.* CNA-PP-422. Alexandria, VA: Center for Naval Analyses. (ADA112792)

Salter, M.S. (1987). *Bradley fighting vehicle gunnery training devices: Trainer attitudes.* RN 87-39. Ft. Benning, GA: U.S. Army Research Institute Field Unit. (ADA183741)

Shlechter, T.M., Kraemer, R.E., Bessemer, D.W., Burnside, B.L., & Anthony, J. (1996). *Perspectives on the virtual training program from members of its initial observer/controller team.* ARI-RR-1691. . Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA310080)

Wiekhorst, L. & Dixon, K.H. (1987). *F-16 limited field of view simulator training effectiveness evaluation: Final report.* Eglin AFB, FL: Tactical Air Command. (ADB115730)

Wiekhorst, L. (1987). *F-16 limited field of view simulator training effectiveness evaluation: Executive summary.* Eglin AFB, FL: Tactical Air Command. (ADB115681)

Reference List A-9. Judgment (Analysts)

Andre, C.R., Wampler, R.L., & Olney, G.W. (1997). *Battle staff training system in support of force XXI training program: Methodology and lessons learned.* RR 1715. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA338728)

Bessemer, D.W. & Myers, W.E. (1998). *Sustaining and improving structured simulation-based training.* RR 1722. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA344895)

Bruce, P.D., Killion, T.H., Rockway, M., & Povenmire, H. (1991). *B-52 and KC-135 mission qualification and continuation training: A review and analysis.* AL-TR-1991-0010. Dayton, OH: University of Dayton Research Institute. (ADA241591)

Kraemer, R.E. & Bessemer, D.W. (1987). *U.S. tank platoon training for the 1987 Canadian army trophy (CAT) competition using a simulation networking (SIMNET) system.* RR 1457. Ft. Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences

Lynn, J. & Palmer, K.L. (1991). *Independent operational assessment of the close combat tactical trainer (CCTT).* OA-0200. Alexandria, VA: U.S. Army Operational Test and Evaluation Command. (ADB160088)

McDade, M.B. (1987). *Training effectiveness analysis (TEA) for the Bradley fighting vehicle system (BFVS), block II modification program.* Ft. Benning, GA: Analysis and Studies Office, DOTD, USAIS. (ADB111660)

Richardson, Bellows, Henry, & Co. Inc. (1951). *Evaluation of the effectiveness of the operational flight trainer F-9-F: Device No. 2-F-13.* Washington, DC: Author. (ADB234094)

Reference List A-10. Analysis (Evaluate)

Ambruster, R.F. (1987). *Training effectiveness evaluation: OH-58D pilot and observer combat skills instruction.* Ft. Rucker, AL: U.S. Army Aviation Center. (ADA190938)

Bailey, S.S. & Hodak, G.W. (1994). *Live-fire versus simulation: A review of the literature.* Special Report 94-002. Orlando, FL: Naval Air Warfare Center Training Systems Division. (ADB189243)

Berg, R.M., Adedeji, A.M., & Trenholm, C. (1993). *Simulation offset to live fire training study: Assessment of Marine Corps live fire training support.* CIM-238. Alexandria, VA: Center for Naval Analyses. (ADB173795)

Burnside, B.L. (1990). *Assessing the capabilities of training simulations: A method and simulation network (SIMNET) application.* RR 1565. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226354)

Drucker, E.H. & Campshure, D.A. (1990). *An analysis of tank platoon operations and their simulation on simulation networking (SIMNET).* RP 90-22. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA226956)

Ehrlich, J.A., Knerr, B.W., Lampton, D.R., & McDonald, D.P. (1997). *Team situational awareness training in virtual environments: Potential capabilities and research issues.* TR 1069. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337606)

Finley, D.L. (1997). *Simulation-based communications realism and platoon training in the Close Combat Tactical Trainer (CCTT).* TR 1064. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA337692)

Fusha, J.E. (1989). *Simulation networking (SIMNET): Evaluation of institutional/ USAIS (U. S. Army Infantry School) use of SIMNET-T. Phases 1 and 2.* RN 2-89. Ft. Benning, GA: U.S. Army Infantry School. (ADA137722)

Hahn, C.P., Krug, R.E., Rosenbaum, H., Stoddart, S.C., & Harmon, J. (1986). *Evaluation of the U.S. Army basic skills education program.* TR 277. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA178650)

Hall, E.R. & Rizzo, W.A. (1975). *An assessment of U.S. Navy tactical team training.* TR 18. Orlando, FL: Training Analysis and Evaluation Group. (ADA011452)

Harman, J. (1984). *Three years of evaluation of the Army's basic skills education program.* RR 1380. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA170476)

Leatherwood, N.J., Schisser, J.S. & Russell, R.J. (1986). *OH-58D aircrew cost and training effectiveness analysis (OH-58D aircrew CTEA): Final report.* TRADOC ACN 85216. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB112520)

Lynn, J. & Palmer, K.L. (1991). *Independent operational assessment of the close combat tactical trainer (CCTT).* OA-0200. Alexandria, VA: U.S. Army Operational Test and Evaluation Command. (ADB160088)

McDade, M.B. (1986). *Bradley fighting vehicle (BFV) training developments study (TDS).* Ft. Benning, GA: Analysis and Studies Office, U.S. Army Infantry School. (ADA173795)

Mitchell, G.W. (1996). *Application of distance learning technology to strategic education.* Carlisle Barracks, PA: U.S. Army War College. (ADA308992)

Morrison, J.E. & Orlansky, J. (1997). *The utility of embedded training.* IDA Document D-1976. Alexandria, VA: Institute for Defense Analyses. (ADA349875)

Operational Research and Analysis Establishment (1990). *Comparison of conventional and simulator enhanced tank gunnery training methods.* ORAE Project Report 523. Ottawa, Canada. Author.

Orlansky, J. & String, J. (1977). *Cost-effectiveness of flight simulators for military training volume I: Use and effectiveness of flight simulators.* IDA Paper P-1275. Alexandria, VA: Institute for Defense Analyses. (ADA052801)

Sassone, P.G., Bercos, J., & Holmgren, J.E. (1986). *Training extension course cost and training effectiveness analysis methodology.* RN 86-14. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA168212)

Shlechter, T.M. (1986). *An examination of the research evidence for computer-based instruction in military training.* TR 722. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA174817)

Simutis, Z.M., Ward, J.S., Harman, J., Farr, B.J., & Kern, R.P. (1988). *ARI research in basic skills education.* RR 1486. Alexandria, VA : U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA201402)

Skog, D., Neal, M.A., & Fields, J.E. (1994). *Breacher cost and training effectiveness analysis (CTEA): Volume I - Main report.* TRAC-WSMR-CTEA-93-019-1. Volume II - Appendices. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB191580) (ADB191753)

Taylor, B.E., Ellis, J.A., & Baldwin, R.L. (1987). *Current status of navy classroom training: A review of 100 navy courses with recommendations for the future.* NPRDC TR 88-11. San Diego, CA: Navy Personnel Research and Development Center. (ADB122161)

Thomas, B.W. & Gainer, C.A. (1990, May). Simulation networking: Low fidelity simulation in U.S. Army aviation. *Proceedings of the Royal Aeronautical Society* (pp. 18.1-18.11). London, England.

White, B.L. & Green, E.K. (1994). *Cost/benefit analysis of video-teletraining for the Marine Corps.* TR 94-008. Orlando, FL: Naval Training Systems Center. (ADB190270)

Winkler, J.D., Shanley, M.G., Crowley, J.C., Madison, R.A., Green, D., Polich, J.M., Steinberg, P., & McDonald, L. (1996). *Assessing the performance of the Army reserve components school system.* MR-590-A. Santa Monica, CA: Rand. (ADA317343)

Reference List A-11. Analysis (Compare)

Ellis, J.A. & Parchman, S. (1994). *The interactive multisensor analysis training (IMAT) system: A formative evaluation in the aviation antisubmarine warfare operator (AW) class "A" school.* NPRDC TN-94-20. San Diego, CA: Navy Personnel Research and Development Center. (ADA285959)

Adkins, S., Belfer, B., Carter, J., Levy, L., Miller, G., Muller, D., Reeded, M.J., Rodriquez, L., & Stinson, D.R. (1988). *The non-line-of-sight weapon system cost and training effectiveness analysis final report.* Washington, D.C.: U.S. Office of Personnel Management. (ADB147263) (Executive summary published separately as ADB147262)

Carroll, D.K. (1995). *Heavy assault bridge cost and training effectiveness analysis (CTEA) final report.* USAES-CTEA-95-HAB-001. Ft. Leonard Wood, MO: U.S. Army Engineer School. (ADA311901)

Crawford, A. & Suchan, J. (1996). *Understanding videoteleducation: An overview.* NPS-SM-96-003. Monterey, CA: U.S. Naval Postgraduate School. (ADA319585)

Green, E.K., McDaniel, R.D., & Dunlap, S.W. (1986). *Economic analysis of heavy duty truck driver trainer alternatives.* Unpublished report. Orlando, FL: Naval Training Systemcs Center.

Keller, A.R., Maruna, R.E., Hawkins, K.A., & Bealieu, H.H. (1991). *Aviation combined arms tactical training training development study (TDS) Phase II.* Report No. AVNC-DOTD-92-1. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB161932)

Lineback, W., Reynolds, M., Everett, J., Gardner, C. Bergman, M., & Stefonek, N. (1983). *Fire support team vehicle cost and training effectiveness analysis (FISTV CTEA) volume 1: Main report.* TRASANA CTEA 15-83. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB955021)

Muller, D., Adkins, S., Belfer, B., Carter, J., & Levy, L. (1988). *The non line of-sight weapon system cost and training effectiveness analysis.* Ft. Bliss, TX: U.S. Army Air Defense Artillery School. (ADB147262)

Noble, J.L. & Johnson D.R. (1991). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 1: Executive summary.* TRAC-WSMR-CTEA-91-018-1. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157064)

Noble, J.L. & Johnson D.R. (1991). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 2: Main report.* TRAC-WSMR-CTEA-91-018-2. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB173567)

Stoloff, P.H. (1991). *Cost-effectiveness of U.S. Navy video teletraining system alternatives.* CRM Research Memorandum 91-165. Alexandria, VA: Center for Naval Analyses. (ADB171320)

Reference List A-12. Analysis (Optimize)

Ellis, J.A. & Parchman, S. (1994). *The interactive multisensor analysis training (IMAT) system: A formative evaluation in the aviation antisubmarine warfare operator (AW) class "A" school.* NPRDC TN-94-20. San Diego, CA: Navy Personnel Research and Development Center. (ADA285959)

Adkins, S., Belfer, B., Carter, J., Levy, L., Miller, G., Muller, D., Reeded, M.J., Rodriquez, L., & Stinson, D.R. (1988). *The non-line-of-sight weapon system cost and training effectiveness analysis final report.* Washington, D.C.: U.S. Office of Personnel Management. (ADB147263) (Executive summary published separately as ADB147262)

Carroll, D.K. (1995). *Heavy assault bridge cost and training effectiveness analysis (CTEA) final report.* USAES-CTEA-95-HAB-001. Ft. Leonard Wood, MO: U.S. Army Engineer School. (ADA311901)

Crawford, A. & Suchan, J. (1996). *Understanding videoteleducation: An overview.* NPS-SM-96-003. Monterey, CA: U.S. Naval Postgraduate School. (ADA319585)

Green, E.K., McDaniel, R.D., & Dunlap, S.W. (1986). *Economic analysis of heavy duty truck driver trainer alternatives.* Unpublished report. Orlando, FL: Naval Training Systemcs Center.

Keller, A.R., Maruna, R.E., Hawkins, K.A., & Bealieu, H.H. (1991). *Aviation combined arms tactical training training development study (TDS) Phase II.* Report No. AVNC-DOTD-92-1. Ft. Rucker, AL: U.S. Army Aviation Center. (ADB161932)

Lineback, W., Reynolds, M., Everett, J., Gardner, C. Bergman, M., & Stefonek, N. (1983). *Fire support team vehicle cost and training effectiveness analysis (FISTV CTEA) volume 1: Main report.* TRASANA CTEA 15-83. White Sands, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB955021)

Muller, D., Adkins, S., Belfer, B., Carter, J., & Levy, L. (1988). *The non line of-sight weapon system cost and training effectiveness analysis.* Ft. Bliss, TX: U.S. Army Air Defense Artillery School. (ADB147262)

Noble, J.L. & Johnson D.R. (1991). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 1: Executive summary.* TRAC-WSMR-CTEA-91-018-1. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB157064)

Noble, J.L. & Johnson D.R. (1991). *Close combat tactical trainer (CCTT) cost and training effectiveness analysis, Volume 2: Main report.* TRAC-WSMR-CTEA-91-018-2. White Sands Missile Range, NM: TRADOC Analysis Command. (ADB173567)

Stoloff, P.H. (1991). Cost-effectiveness of U.S. Navy video teletraining system alternatives. CRM Research Memorandum 91-165. Alexandria, VA: Center for Naval Analyses. (ADB171320)

Reference List A-13. Survey

Bretl, D.C., Rivera, F.G., & Coffey, B.A. (1996). *Engineer combined arms tactical trainer (ENCATT) cost and training effectiveness analysis (CTEA).* TRAC-WSMR-CTEA-96-008. White Sands Missile Range, NM: U.S. Army TRADOC Systems Analysis Activity. (ADB213104)

Brown, F.J. (1978). *The Army training study: Administration.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186321)

Brown, F.J. (1978). *The Army training study: Training effectiveness analysis, volume I: Armor.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186323)

Brown, F.J. (1978). *The Army training study: Training effectiveness analysis, volume II: Armor.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186324)

Brown, F.J. (1978). *The Army training study: Training effectiveness analysis, volume IV: Ordnance, signal, and computer assisted map maneuver system (CAMMS).* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186326)

Brown, F.J. (1978). *The Army training study: Training effectiveness analysis: Summary.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA186322)

Brown, F.J. (1978). *The Army training study: Data book.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA184393)

Brown, F.J. (1978). *The Army training study: Report summary.* Ft. Monroe, VA: U.S. Army Training and Doctrine Command. (ADA184392)

Fusha, J.E., Renn, A.N., & Thompson, T.J. (1984). *Training effectiveness analysis: Status of institutional and unit mortar training.* RR 1367, Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (ADA158018)

George, E., Jackson, G., Kenney, M., & Kilgore, E. (1991). *Combat service support (CSS) standard Army management information systems (STAMIS) tactical Army combat service support computer system (TACCS) training effectiveness analysis (TEA) update.* White Sands Missile Range, NM: TRADOC Analysis Command. (ADB166701)

Hall, E.R. & Rizzo, W.A. (1975). *An assessment of U.S. Navy tactical team training.* TR 18. Orlando, FL: Training Analysis and Evaluation Group. (ADA011452)

Pugh, H.L., Parchman, S.W., & Simpson, H. (1991). *Field survey of videoteletraining systems in public education, industry, and the military.* NPRDC TR-91-7. San Diego, CA: Navy Personnel Research and Development Center. (ADA234875)

Rosenblum, D.E. (1979). *Combat effective training management study (CETRM).* Washington, DC: CETRM. (ADA101993)

Walsh, W.J., Gibson, E.G., Miller, T.M., Hsieh, P.Y., Gettman, D., & Newcomb, S. (1996). *Characteristics of distance learning in academia, business, and government.* AL/HR-TR-1996-0012. Brooks AFB, TX: Human Resources Directorate - Technical Training Research Division. (ADA311369)

APPENDIX B. ACRONYMS

AAR—After-Action Review
AFHRL—Air Force Human Resources Laboratory
ALSP—Aggregate Level Simulation Protocol
AMTEP—ARTEP Mission Training Plan
ARI—U.S. Army Research Institute for the Behavioral and Social Sciences
ARPA—Advanced Research Projects Agency (formerly DARPA)
ARTEP—Army Test and Evaluation Program
ARTS—Army Training Study (model)
ASW—Anti-Submarine Warfare
BBS—Brigade/Battalion Battle Simulation
BFTT—Battle Force Tactical Trainer
BFV—Bradley Fighting Vehicle
CA—Cost Analysis
CAS—Close Air Support
CAT—Canadian Army Trophy
CATT—Combined Arms Tactical Trainer
CBA—Cost-Benefit Analysis
CBI—Computer-Based Instruction
CBP—Comparison-Based Prediction
CBS—Corps Battle Simulation
CCTT—Close Combat Tactical Trainer
CEA—Cost-Effectiveness Analysis
CEAT—Cost-Effectiveness Analysis of Training
CEV—Combat Engineering Vehicle
CNA—Center for Naval Analyses
CNET—Chief of Naval Education and Training
COEA—Cost and Operational Effectiveness Analysis
COFT—Conduct of Fire Trainer
CTEA—Cost and Training Effectiveness Analysis
DARPA—Defense Advanced Research Projects Agency (now ARPA)
DEOMI—Defense Equal Opportunity Management Institute
DFO/MULE—Deployed Forward Observer/Modular Unit Laser Equipment
DIS—Distributed Interactive Simulation
DMDC—Defense Manpower Data Center
DMSO—Defense Modeling and Simulation Office
DoD—Department of Defense
DTIC—Defense Technical Information Center
ENCATT—Engineer Combined Arms Tactical Trainer
GUARDFIST—Guard Unit Armory Device, Full Crew Interactive Simulation Trainer
HUMRRO—Human Resources Research Organization
ICAI—Intelligent Computer-Aided Instruction—
IDA—Institute for Defense Analyses
IMAT—Interactive Multisensor Analysis Training
ISD—Instructional Systems Development
ITS—Intelligent Tutoring System
ITV—Instructional TV
IVD—Interactive Video Disk
JSIMS—Joint Simulation System

JTCTS—Joint Tactical Combat Training System
JTIDS—Joint Tactical Information Display System
JTTP—Joint Tactics, Techniques, and Procedures
LM—Learning Methodology
LMWG—Learning Methodology Working Group
LSTS—Large-Scale Training Simulations
M&S—Models and Simulations
MAIS—Major Automated Information System
MARSIM—Maritime Simulation
MCOFT—Mobile Conduct of Fire Trainer
MDAP—Major Defense Acquisition Program
MDT2—Multi-service Distributed Training Testbed
METT-T—Mission, Enemy forces, Troops friendly, Terrain control, Time
MIL-STD—Military Standard
MOE—Measure of Effectiveness
MOP—Measure of Performance
MOS—Military Occupational Specialty
MPT—Manpower, Personnel, and Training
NAWC—Naval Air Warfare Center
NAWCTSD—Naval Air Warfare Center Training System Division
NLOS—Non Line of Sight
NPRDC—Navy Personnel Research and Development Center
NTC—National Training Center
O/C—Observer/Controller
ODUSD(R)—Office of the Deputy Under Secretary of Defense for Readiness
OOTW—Operations Other Than War
OPTEMPO—Operating Tempo
ORD—Operational Requirements Document
OSD—Office of the Secretary of Defense
PGT—Platoon Gunnery Trainer
PGTS—Precision Gunnery Training System
POI—Program of Instruction
R&D—Research and Development
SAT—Systems Approach to Training
SIMCAT—Simulation in Combined Arms Training
SIMNET—Simulator Network
SME—Subject-Matter Expert
STOW—Synthetic Theater of War
STOW ACTD—Synthetic Theater of War Advanced Concept Technology Demonstration
STRICOM—U.S. Army Simulation, Training and Instrumentation Command
TADSS—Training Aids, Devices, Simulators, and Simulations
TAEG—Training Analysis and Evaluation Group
TARGETs—Targeted Acceptable Response To Generated Events Or Tasks
TCEF—Training and Cost-Effectiveness File
TD—Training Device
TD/S—Training Device/Simulator
TDR—Training Device Requirement
TEA—Training Effectiveness Analysis (or Training Effectiveness Assessment)
TECATS—Training Effectiveness Catalog System
TEXCOM—Test and Experimentation Command
TOM—Teamwork Observation Measure
TQM—Total Quality Management

TRAC—TRADOC Analysis Center
TRADOC—Training and Doctrine Command
UJTL—Universal Joint Task List
UPAS—Unit Performance Assessment System
USD(P&R)—Under Secretary of Defense for Personnel and Readiness
VIGS—M1 Videodisc Interactive Gunnery Simulator
VTT—Video Teletraining
VV&A—Verification, Validation, and Accreditation
WARSIM—Warfighter's Simulation
WSAP—Weapon System Acquisition Process

AUTHOR INDEX

Aagard, J.A., 93, 94
Ackerson, G., 16
Adedeji, A.M., 118, 128
Adkins, S., 44
Alluisi, E.A., 50, 51, 78, 80, 107, 108, 124
Ambruster, R.F., 121
Ames, W., 16
Anderson, R.D., 16
Angier, B.N., 51
Anthony, J., 54, 55, 102, 119
Aylward, M., 51, 120, 121
Babbitt, B.A., 91, 92
Bailey, M., 51, 120, 121
Baker, J.D., 74, 75, 77
Bangert-Drowns, R.L., 16
Bauer, R.W., 123
Bealieu, H.H., 88, 89, 128
Belfer, B., 44
Bell, H.H., 39, 58, 65, 66, 78, 80, 107, 108
Bell, S.A, 28
Berg, R.M., 118, 128
Berger, S.C., 53, 54, 57, 102
Bessemer, D.W., 18, 31, 35, 39, 52, 53, 54, 55, 57, 62, 81, 86, 87, 92, 96, 102, 115, 119, 125, 130
Blackwell, N.J., 123
Bloom, B. S., 16
Boldovici, J.A., 35, 52, 81, 86, 87, 92, 107, 115, 124
Bouchard, T.J., 47, 73, 91
Bowne, D.S., 35
Boycan, G.G., 123
Braby, R., 121
Bretl, D.C., 116
Brown, F.J., 47, 48
Brown, J.B., 29
Brown, R., 54, 57, 102
Brown, R.E., 27, 52, 54, 102
Browning, R.F., 32, 33, 34, 86, 87
Burke, E., 51
Burnside, B.L., 45, 52, 53, 54, 56, 62, 96, 102, 118, 119
Butler, W.A., 46, 90
Calvin, J., 51
Campbell, C.H., 74
Campbell, D.T., 18, 25, 26, 27, 28, 29, 30, 33, 61, 74, 75, 85, 86, 87
Campbell, R.C., 74
Campshure, D.A., 45, 52, 54, 56, 102, 118, 125
Cannon-Bowers, J.A., 96
Carroll, D.K., 43, 44
Carter, J., 44
Chatterjee, A., 51
Clapper, D., 52
Coffey, B.A., 116
Cohen, J., 85, 86
Cohen, P., 85, 86
Cohen, P.A., 16
Colburn, E., 66, 78, 79
Collins, J.W., 119
Communication Technology Applications, Inc., 42
Cook, T.D., 25, 86, 87
Cosby, N.L., 51
Crane, P.M., 53, 54, 57, 102
Crawford, A.M., 44
Cross, K.D., 32, 54, 55, 102, 119
Dahlman, C.J., 51, 121
Davis, M.S., 36, 120, 121, 125
Department of Defense, 11, 12, 22, 104, 105
Department of Defense, Office of the Inspector General, 2, 3
Department of the Army, 11
Derrick, D.L., 36, 120, 121, 125
Djang, P.A., 46, 54, 56, 90, 102, 128
Drucker, E.H., 45, 52, 54, 56, 102, 118, 125
Dwyer, D.J., 58, 61, 63, 65, 78, 80, 96, 103, 107, 108
Dyer, J.L., 95
Ellis, J.A., 42, 129
Ennis, J.J., 32
Evans, R.M, 46, 121
Fallesen, J.J., 80, 96, 107, 108
Farr, B.J., 42
Farrow, S., 66, 78, 79
Fields, J.E., 43
Fingerman, F.W., 123
Finley, D.L., 46, 54, 56, 89, 102, 118
Fish, D., 51, 120, 121
Fitz-Gibbon, C.T., 91, 92, 93, 94
Fletcher, J.D., 15, 54, 58, 102, 127
Flynn, M.R., 74
Fober, G.W., 95
Fowler, F.J., 47, 91
Fowlkes, J.E., 61, 63, 96, 103
Franz, T., 61, 96, 103
Fusha, J.E., 45, 54, 55, 102
Gage, N.L., 25, 86
Gainer, C.A., 45, 53, 54, 56, 102, 119
Gardner, C.V., 32
Garlinger, D.K., 80, 96, 107, 108
Gawlick-Grendell, L.A., 35

George, E., 129
Glaser, R., 14
Gleisner, D., 2
Gound, D., 52, 53, 54, 102
Graham, S.E., 54, 102
Greene, D.E., 27
Gurwitz, R., 51
Hagman, J.D., 35
Hall, E.R., 46, 48, 93, 94
Hammon, C.P., 51, 121
Harman, J., 28, 42, 46
Harris, K., 28
Harris, M., 51
Harrison, J.A., 39
Hart, R.J., 35
Hartley, D.S., 53, 54, 55, 102
Hawkins, K.A.,, 88, 89, 128
Haynes, M.D., 27
Henerson, M.E., 91, 92
Herman, J.L., 93, 94
Hiller, J.H., 31, 52, 74, 80, 95, 107, 108, 124
Hodak, G.W., 121
Hoffman, G.R., 53, 54, 57, 102
Hoffman, R.G., 86, 87
Hofmann, D.A., 96
Holstead, J.R., 53, 54, 57, 102
Holz, R.F., 95
Honig, J.G., 31, 58, 60, 62, 63, 64, 65, 119, 120
Horey, J.D., 31, 46, 87, 89, 90
Horowitz, S.A., 51
Houck, M.R., 39
Houser, T.L., 96
Hughes, C.R., 46, 90
Jackson, G., 129
Jeantheau, G.G., 17, 20, 38, 94
Johns, P., 46, 85, 90
Johnson D.R., 43, 54, 56, 102, 118
Johnson, E.M., 74, 75, 77
Johnson, J.L., 127
Jones, S., 31
Kaempf, G.L., 33, 34, 123
Kaminski, P.G., 3
Kass, R., 87, 88
Keller, A.R., 39, 88, 89, 128
Kelly, J.F., 39, 117, 118
Kenney, M., 129
Kilgore, E., 129
Kirkpatrick, D.L., 93, 94, 106
Klein, G.A., 46, 85, 90
Kolasinski, E. M., 52
Kolosh, K.P., 54, 55, 102
Kraemer, R.E., 39, 52, 54, 57, 81, 82, 83, 102, 115, 119
Kribs, H.D., 15
Kruse, K.L., 53, 54, 55, 102
Kulik, C.C., 16
Kulik, C.-L.C., 16
Kulik, J.A, 16
Laferriere, R., 46, 54, 56, 90, 102, 128
Lampton, D.R., 28
Lane, N.E., 61, 96, 103
Laughy, N., 28
Learning Methodology Working Group, 114, 130
Leatherwood, N.J., 118
Leibrecht, B.C., 35, 107, 124
LeMaster, W.D., 123
Leonard, R.L, 123
Levine, D.B., 31, 58, 60, 62, 63, 64, 65, 119, 120
Levy, L., 44
Lewman, T., 95
Lickteig, C.W., 119
Litwin, 91
Loral Systems, 66
Love, J.F., 58, 65, 78, 80, 107, 108
Luiten, J., 16
Lussier, J.W., 78, 80
Lynn, J., 43, 54, 56, 102, 118
Lysakowski, R.S., 16
Macklin, L., 39
Marcus, A.J., 39
Mark, L.J., 15
Martin, M.G., 29
Maruna, R.E., 88, 89, 128
McDade, M.B., 42, 43, 116
McDaniel, W.C., 34
McDonald, R.G., 127
McDonough, J., 66, 78, 79
McFadden, R.W., 123
McFann, H.H., 95
Meister, D., 93, 94
Meliza, L.L., 53, 58, 62, 65, 78, 80, 96, 107, 108
Merriam-Webster, 14, 40
Metzko, J., 51, 121
Mirabella, A., 39, 46, 58, 62, 65, 78, 80, 85, 90, 103, 106, 107, 108, 126, 127
Morris, L.L., 91, 92, 93, 94
Morrison, J., 39, 58, 62, 65, 86, 87, 103, 125, 126, 127
Moses, F.L., 51, 58, 60, 65, 78, 80, 107, 108, 120, 121
Muckler, F.A., 46, 89
Mullen, W.J., 95
Muller, D., 44
Mullis, C., 54, 57, 102
Myers, W.E., 53, 54, 57, 74, 102
Neal, M.A., 43, 129
Nesselroade, P., 54, 55, 102, 119
Noble, J.L., 43, 54, 56, 102, 118
Nystrom, C.O., 91, 92

Orlansky, J., 3, 31, 51, 58, 60, 62, 63, 64, 65, 119, 120, 121, 125
Oser, R.L., 61, 63, 96, 103
Palmer, K.L., 43, 54, 56, 102, 118
Parchman, S.W., 42, 48, 129
Parrish, J.J., 39
Paschal, R., 16
Patton, M.Q., 91, 92
Perez, R., 46, 85, 90
Pfeiffer, M.G., 32, 33, 46, 86, 87, 89, 90
Pishel, R.G., 27, 52, 54, 102, 129
Pleban, R.J., 29
Povenmire, H.K., 32, 34
Psotka, J., 15
Pugh, H. L., 27, 48
Quester, A., 39
Quillinan, J.D., 53, 54, 55, 102
Rabideau, G.F., 93
Rakolta, M.J., 66
Rankin, W.C., 93, 94
Redfield, D.L., 16
Regian, J. W., 76
Resource Consultants, Inc., 5
Rivera, F.G., 116
Rizzo, W.A., 46, 48
Root, J., 95
Roscoe, S.N., 32, 34
Rose, A.M., 123
Rosenblum, D.E., 47
Rousseau, E.W., 16
Rowatt, W.C., 81, 82, 83, 115
Rozen, U., 128
Russell, R.J., 118
Russell, T.L., 81, 83
Salas, E., 61, 96, 103
Salter, M.S., 95, 129
Sanders, J.J., 74
Sassone, P.G., 3
Schaffer, W.A., 3
Schisser, J.S., 118
Schwab, J., 52, 53, 54, 102
Schwalb, B.J., 16
Scott, B.B., 54, 56, 102, 128
Scott, P.G., 34
Semple, C.A., 93, 94
Seymour, G.E., 115
Sheppard, D.J., 127
Sheppe, M.L., 127
Shlechter, T.M., 53, 54, 55, 62, 96, 102, 119
Shute, V.J., 15, 35, 76
Simpson, H., 2, 3, 5, 27, 48, 51, 75, 115, 120, 121
Simutis, Z.M., 42
Skog, D., 43
Slavin, R..E., 16
Smith, B.W., 32, 54, 55, 102, 119
Smith, J.W., 52
Smith, M.D., 35
Smith, S.E., 54, 102
Smode, A.F., 34
Solick, R.E., 78, 80
Solomon, H., 124
Southard L.D., 27, 52, 54, 102
Spector, 91, 92
Stanley, J.C., 18, 25, 26, 27, 28, 29, 30, 33, 61, 85, 86
Stapp, K.M., 129
Steadman, G.W., 128
Sterling, B., 36
Stewart, J.E., 122, 123
Sticha, P., 39, 58, 62, 65, 103, 126, 127
Stoloff, P.H., 42, 43
String, J., 125
Suchan, J., 44
Tan, S.C., 53, 62, 96
Taylor, H.L., 31, 51, 58, 60, 62, 63, 64, 65, 119, 120, 121
TEXCOM, 52, 54, 55, 102, 129
Thomas, B.W., 45, 53, 54, 56, 102, 119
Thomas, G.S., 39
Thorpe, J.A., 123
Trenholm, C., 118
Turnage, J.J., 96
Universal Cities Studios, Inc., 114
Varney, N.C., 123
Walberg, H.J., 16
Ward, J.S., 42
Watson, B.L., 53, 54, 55, 102
Weinstein, T., 16
West, W.D., 2
Wetzel, C. D., 27, 115
Wheaton, G.R., 123
Whitten, T., 31
Willett, J.B., 16
Witmer, B.G., 123
Wood, M.A., 129
Worley, R.D., 51, 120, 121
Yamashita, J.J., 16
Youngblut, C., 51, 121

SUBJECT INDEX

acquisition regulations, DoD, 22-24, 104-105
advanced distributed simulation, 2
after-action review, 21, 53, 60
analogy, 43-44
analysts, 37, 38, 39, 40, 41
analytical evaluation
  definition, 40
  procedural guidance for, 89-90
  relation to judgment-based evaluation, 40
  relative usage, 19
  SIMNET/CCTT examples, 54, 55-56
  strategies, 40, 43-46
attitudes, 37
backward transfer, 33
case studies, evaluation
  LSTS, 49-66
  MDT2, 58-66, 102-103
  methods, media, programs, and simulations, 25-48, 116-129
  SIMNET/CCTT, 50-66, 101-103, 107
collective training, 12, 13, 95-96
comparative evaluation, 17, 38, 41, 47
comparison-based prediction, 90
computer-based instruction, 15
constructive simulation, 2
cost
  cost-effectiveness analysis of training, 2-3
  cost-effectiveness of laboratory versus field evaluations, 75
  cost-effectiveness, 2
  MDT2 cost analysis, 65
  training and cost-effectiveness file, 5-6
cost-effectiveness, 2
criteria, evaluation
  definition, 21-22, 101
  deriving a set for LSTS evaluation, 106-109
  evaluation perspective influence on, 16, 104-105
  expert recommendations, 107-108
  Kirkpatrick's framework, 106-107
  recommendations for LSTS evaluations, 108-109
  See also variables, dependent
data
  archival, 52, 124
  attitudes, 37
  historical, 31, 35, 40-43, 45-46
  internal versus external validity, 26, 75-76
  judgment, 37
  laboratory versus field tests, 75-76
  quality, 115
  technical estimates, 37
data bases, training evaluation
  training and cost-effectiveness file, 5-6
  training effectiveness catalogue system, 5
DEFT, 90
dependent variables
  behavior, 106, 107, 108-109
  collective performance, 22, 106, 108-109
  combat performance, 21
  developer discretion in use of, 104-105
  DoD acquisition regulations governing, 104-105
  evaluation method effects on, 101-102
  for small- versus large-scale evaluations, 102-103
  Kirkpatrick's framework, 106-107
  MDT2 evaluation, 61-64, 102-103
  measures of effectiveness, 104
  measures of performance, 104
  outcome measures, 62, 64
  process measures, 61, 63
  reactions, 21, 103, 106
  results, 103, 107
  SIMNET/CCTT evaluations, 101-103
  student learning, 21-22, 106
  subjective versus objective, 25
  TARGETs measures, 61, 63, 103
  TOMs measures, 61, 62, 64, 103
  See also criteria, evaluation
development of LSTS
  as process versus event, 24, 75
  developmental phases, 23
  DoD directives and regulations governing, 22-24, 104-105
  evaluation objectives, 112
  milestone decision points, 23-24
evaluation
  criteria, 101-109
  dependent variables. *See* dependent variables
  framework. *See* framework, LSTS evaluation
  laboratory versus field, 16, 73-77
  levels, 17
  methods. *See* methods, evaluation
  perspective. *See* perspective, evaluation
  principles. *See* principles, evaluation
  problem areas. *See* problem areas, evaluation

procedural guidance. *See* procedural guidance, evaluation
questions. *See* questions, evaluation
stakeholders, 113
timing. *See* timing, evaluation
case studies. *See* case studies, evaluation
comparative versus non-comparative, 17, 38, 41, 47
ex post facto experiments, 35-36, 124
experiments
definition, 25
ex post facto, 35-36, 124
pre-, 27-29
procedural guidance for, 85-88
quasi-, 29-31
relative usage, 19
SIMNET/CCTT evaluations, 53-55
test, 31-32, 74
transfer, 32-35
transfer, backwards, 33
true, 26-27
expert recommendations, 107-108
extrapolation, 44-45
field evaluation
critiques, 80-84
laboratory versus field evaluations, 16, 73-77
lessons learned in, 77-80
flaws in training research, 81-84
FORTE, 90
framework, LSTS evaluation
building, 9-24, 111-130
definition, 9, 111
framework, system and program evaluation, 93-95
historical data
in analytical evaluations, 40, 40-43, 45-46
in archival analysis, 52, 104
in comparative versus non-comparative evaluations, 41
in ex post facto analysis, 31, 35
in reviews and meta-analyses, 45-46
historical data, 45-46
human factors, 93, 94
hypothetical ways to train, evaluating, 40
individual training, 12-13
inspector general, DoD office of 1997 audit, 2, 3
intelligent computer-aided instruction 15
intelligent tutoring systems 15
joint training, 12-13
judgment-based evaluation
comparative versus non-comparative, 38
data collection methods, 37
definition, 36
estimating training potential, 39-40
on hypothetical ways to train, 36
procedural guidance, 91-93
relative usage, 19
respondent category, 37, 38
similarity to experiment, 36-37
SIMNET/CCTT evaluations, 54, 57
Kirkpatrick's framework, 106-107
laboratory evaluations, 16, 73-77
laboratory versus field evaluations, 16, 73-77
changes in developing system as a factor in, 75
characteristics and difficulties, 73-74
constraints of field evaluations, 74
internal and external validity of data, 75-76
LSTS complexity as a factor in, 74
relevance and cost-effectiveness, 75
learning curves, 29, 63, 66, 115
lessons learned
in field evaluations, 77-80
in MDT2 case study, 65-66, 78-80
in SIMNET/CCTT development, 78-80
levels of evaluation, 17
live simulation, 2
LSTS
case studies, 49-66
complexity, 74
definition, 2
development. *See* development of LSTS
evaluation framework. *See* framework, LSTS evaluation
MDT2
case study, 58-66, 102-103
cost analysis, 65
lessons learned, 65-66, 78-80
measures of effectiveness, 104
measures of performance, 104
media, training, 14-15
method, training, 14-16
methods, evaluation
analytical. *See* analytical evaluation
evaluation criteria relations, 101-102
evaluation levels commonly associated with, 20-21
experiments. *See* experiments
factors influencing selection, 19
frequency of usage, 19-20
judgment-based. *See* judgment-based evaluation
taxonomy, 18
timing, effects on methods used, 19-20, 22-24
used in SIMNET/CCTT, 51-52
survey. *See* survey
milestone requirements, 11, 23-24
modeling, 43

no significant difference finding, 81, 83-84
non-comparative evaluation, 17, 38, 41, 47
objectives, evaluation
- defining, 111-113
- general, 10-11
- stage of system development, effects, 111-112
- 1. estimate need for new training system, 112, 116
- 2A. predict training effectiveness, 112, 117, 118
- 2B. measure training effectiveness, 112, 117, 119-120
- 2C. determining training effectiveness, 112, 117, 120-121
- 3A. predict transfer of training, 112, 122
- 3B. measure transfer of training, 112, 122-125
- 3C. determine transfer of training, 112, 125
- 4A predict user acceptance, 112, 126
- 4B. measure user acceptance, 112, 126-127
- 4C. determine user acceptance, 112, 127
- 5AB. support training design, 112, 127-128
- 5C. determine training status, 112, 127, 129
- 6AB. support system design, 112, 130
- 6C. evaluate system design, 112, 113, 130

outcome measures, 62, 64
perspective, evaluation, 16, 104-105
phases, system development, 23
pre-experiment, 27-29
principles, evaluation
- data quality, obtaining the best, 115
- evaluation as process versus event, 113
- influencing system design and development, 114
- learning curves, developing, 115
- multi-dimensionality of evaluation, 114
- objectives, defining, 113
- stakeholders, determining, 113
- transfer of training, measuring, 115

problem areas, evaluation, 73-84
procedural guidance, evaluation
- analytical, 89-90
- collective, 95-96
- experiment, 85-88
- judgment and survey, 91-93
- system and program evaluation frameworks, 93-95, 106-107
- team training, 95, 96

process measures, 61, 63
program, training
- definition, 14
- description, 15
- frequency of evaluations, 15
- SIMNET program for armor battalions, 74

qualitative evaluation, 17
quasi-experiment, 29-31
questions, evaluation
- how to evaluate?, 16
- overview, 9-10
- what are evaluation criteria?, 21-22
- what is evaluated?, 14-16
- when to evaluate?, 22-24
- where to evaluate?, 16
- whose training is evaluated?, 12-13
- why to evaluate?, 10-11

reaction measures, 21, 103, 106
results measures, 103, 107
SIMNET/CCTT
- case study, 50-66, 101-103, 107
- cost-effectiveness, 2
- lessons learned, 78-80
- training program for armor battalions, 74

simulated transfer, 87
simulation, training
- definition, 14
- description, 14-15
- frequency of evaluations, 15

stakeholders, evaluation, 113
stand-alone, single-system simulation, 2
strategies, analytical evaluation, 40, 43-46
student learning measures, 21-22, 106
subject-matter experts, 37
subjective versus objective measures, 25
survey
- comparative versus non-comparative, 47
- data collection methods, 47
- definition, 47
- MDT2 evaluation, 62, 65
- procedural guidance for, 91-93
- purposes, 47
- relative usage, 19
- SIMNET/CCTT evaluations, 54, 58
- small- versus large-scale, 47-48

system design, assessment, 112, 113, 114, 130
TARGETs, 61, 63
task list analysis, 45
team training, 12, 13, 95-96
technical estimates, 37
test, 31-32, 74
timing, evaluation
- developmental phases, 23
- DoD directives and regulations governing, 22-24, 104-105
- effect on evaluation methods used, 19-20, 22-24
- milestone decision points, 23-24

TOMs, 61, 62, 64
total quality management, 11
training

- content, 13
- echelon, 12-13
- individual versus collective, 12
- method of development, 4-7
- organization, 1, 7
- task lists, defining, 106
- versus education, 13

training and cost-effectiveness file, 5-6
training design, assessment, 112, 127-129
training effectiveness analyses, 3
training effectiveness catalogue system, 5
training mix model, 91
training research, common flaws in

- ceiling effect, 82
- device system errors, 82
- floor effect, 83
- groups treated differently, 82
- insufficient practice, 83
- small sample size, 82
- subjects not random or matched, 82
- unreliable performance measures, 82

transfer effectiveness ratios, 32
transfer of training

- backward transfer, 33
- collective performance, 124
- definition, 17, 18, 124
- evaluation objective, 112, 122, 123-125
- evaluation principle, 115
- experiments, 32-35
- measuring during evaluation, 115
- simulated, 87

true experiment, 26-27
unit performance assessment system, 53, 62
universal joint task list, 106
UPAS, 53, 62
user acceptance, 37, 112, 126-127
validity. *See* data
variables, dependent. *See* dependent variables
verification, validation, and accreditation, 105
virtual simulation, 2